

THE EFFECT OF CHANGED MATERIAL ON ABILITY TO DO FORMAL SYLLOGISTIC REASONING

BY
MINNA CHEVES WILKINS

Submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy in the Faculty of Philosophy,
Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R S WOODWORTH, Editor

No. 102

NEW YORK
November, 1928

ACKNOWLEDGMENTS

The author is indebted to Dr. L. L. Thurstone for suggesting the field from which the problem was chosen; to Professor R. S. Woodworth for his unfailing interest and many valuable suggestions during the progress of the experiment and to Professor A. T. Poffenberger for assistance and encouragement at a crucial point of the research.

TABLE OF CONTENTS

	PAGE
CHAPTER I	
1. Statement of problem	5
2. Preliminary experimentation	8
CHAPTER II—Main experiment ..	12
1. Method	12
2. Results	22
Difference in difficulty caused by change in material.	
a. Correlation between different parts of the test	23
b. Mean scores	24
c. Distribution of scores	24
3. Conclusions ...	25
4. Effect of change in material on correlation with intelligence test	28
CHAPTER III—Difficulty of each item and each kind of fallacy as affected by change of material	33
1. Discussion	33
2. Tables	35
a. Items grouped according to kind of fallacy involved, and with each item a measure of its difficulty, its diagnostic value for the syllogism test and its diagnostic value for the intelligence test	35
b. Average measures obtained from above for each kind of fallacy	44
c. Syllogisms used in the test as each appeared in the four different kinds of material. Measures of difficulty, etc., given with each	45
3. Conclusions	68
CHAPTER IV—Summary of conclusions	77

CHAPTER I

1. STATEMENT OF THE PROBLEM

In attempting to teach logic to college students one is struck with their naive attitude toward exact thinking. To many of them to determine the exact meaning of a statement seems to be a thoroughly new task. They feel that a hazy notion of the content of the proposition is all that could be demanded of them. Many of them are not only ignorant of any guides to thinking but are unconscious of a need for exact and careful thinking. Woodworth¹⁴ says, "The reasoner needs a clear and steady eye in order to see the conclusion that is implicated in the premises. Without this he falls into confusion and fallacy or fails, with the premises both before him, to get the conclusion. The 'clear and steady mental eye' in less figurative language means the ability to check hasty responses to either premise alone or to the extraneous features of the situation so as to insure that unitary response to the combination of premises which constitutes the perceptive act of influence." There are many college students who possess this "clear and steady mental eye" only on occasions and are very easily distracted by "extraneous features of the situation." Unfortunately the college student is not alone in this characteristic but shares it with the rest of humanity.

A college class in logic seems to divide itself very soon into two groups, those who seize on the material given at once and in a few hours understand everything that it takes weeks for the other group to assimilate, if it ever does assimilate them. This second group seems to understand principles fairly well as long as they are applied to facts within their experience, but seems unable to apply any principles as soon as the material is unfamiliar or symbolic. For instance, they see clearly that if you have the proposition, "All horses are animals," you cannot logically deduce from that the proposition that, "All animals are horses"; but whenever the proposition is, "All x's are y's," they feel sure that this necessarily implies, "All y's are x's." Many of the members of this second group stand above the median in most of their classes. Some of them have unusual verbal facility. One student who failed to make a pass-

ing grade in a logic course won an intercollegiate debate. She was eloquent and was an adept at arousing the emotion of her hearers. The judges were but human. These vague questions and impressions were crystalized into the following problems.

1. The main problem was to determine to what extent ability to do formal syllogistic reasoning is affected by changing the material reasoned about. Does ability to do this sort of reasoning when the terms are concrete familiar material mean necessarily equal ability when the terms are symbols or long unfamiliar words? How much, if any, will this ability be affected by making the material deliberately distracting or suggestive?

2. If ability to do formal syllogistic reasoning in familiar material is not identical with ability to do this sort of reasoning with other material, which of these abilities will have the closest relationship to general intelligence as measured by an intelligence test?

3. Which sorts of fallacies will be most difficult in different types of material?

4. Which fallacies will have most value in determining success in a syllogism test of this kind?

5. Success in which fallacies will have the highest correlation with success in a test for general intelligence?

Thorndike¹¹ reports* an experiment on the effect of changed data on reasoning. He had 97 graduate students solve a series of nine simple problems in algebra, the problems being expressed in the usual way and then changed to a less usual expression. The amount of change varied. He concludes that "any disturbance whatsoever in the concrete particulars reasoned about will interfere somewhat with the reasoning, making it less correct or slower or both." He cites these results as indicating the importance of habit in reasoning, and defines reasoning as the organization and cooperation of habits.

Bailor² in a recent study gives data on the effect of content and form in tests of intelligence. His data did not allow of a determination of the effect of changed content on difficulty but yield correlations between standings in groups of tests where the form is the same and the content varied. He defines form as the objective mold in which the test is cast; the objective arrangement in which the problem is presented; for example, completion tests, analogies tests. Content he defines as kinds of data which will be limited generally to three

* The references are found on page 79

types: (1) verbal, (2) numerical, and (3) spatial. The original data for this study were obtained by Thorndike in his study of Mental Discipline in High School Studies. Batteries of varied tests were given to 1039 school children and were repeated after a year's interval. Bailor grouped the tests according as they were (1) word-meaning or verbal tests, (2) tests primarily of numerical relations and (3) tests of spatial or geometric relations. The mean reliability coefficients for these groups ranged from .608 to .766. The mean intercorrelations were (1) words and (2) numbers, .5604; (2) numbers and (3) space, .508; (3) space and (1) words, .46. He also separated out all the tests of one particular form, such as analogies tests and computed the correlation between standings in grammatical analogies, spatial analogies and verbal analogies. These mean coefficients ranged from .33 to .39 and the mean reliability coefficients ranged from .484 to .656. In generalization tests he found a mean correlation of .213 between verbal and numerical material where the mean reliability coefficient for the verbal material was .5068 and for the numerical .24. In completion tests he found a correlation of .149 between verbal and numerical material. For a group of 184 cases he gives correlations between the Terman Group Intelligence Test and the respective content groups. The "words" group had a correlation of .728 with the Terman, the "numbers" group a correlation of .604 and the "space" group a correlation of .59, the "words" group apparently measuring more nearly than the other groups whatever is measured by the intelligence test. Bailor concludes that the correlations throughout are positive, and that differences in the relative standings of pupils occur when they are given tests having differences either in form or content.

The present study relates itself more nearly to the Thorndike¹¹ study of the effect of changed data on reasoning than to this study of Bailor's, in that one particular type of test is used, and the change in difficulty caused by change of material is studied as well as change in relative standing. In this study the form is kept more exactly constant as the material is changed. Each item has its counterpart in form in all parts of the test.

2. PRELIMINARY EXPERIMENTATION

The first work was done at Carnegie Institute of Technology where the first arrangement of the syllogism test was given to 163 second year students of the Margaret Morrison School, the School of Industries, and the School of Engineering. The form of the test then used consisted of 128 items, each item consisting of two premises and a conclusion, or, in a few cases, of one proposition with a conclusion deduced therefrom. Thirty-two of these items were expressed in familiar material, thirty-two in symbolic terms (a's and b's or x's and y's), thirty-two in unfamiliar material (long and unfamiliar words), and thirty-two in what might be called suggestive material (familiar and concrete terms with the meaning designed to be misleading). The items of different material were arranged in apparently random order. On each page of the test, the first, fifth, and ninth items were of familiar material; the second, eighth, and tenth were of symbolic material; the third, sixth, and eleventh were of unfamiliar material; and the fourth, seventh, and twelfth were of suggestive material. The subjects were allowed to work on the test for forty-five minutes, exclusive of the time taken in reading directions. The different material was scored separately with the following results in mean score. The score was per cent right of those attempted.

<i>School</i>	<i>Number</i>	<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Total</i>
Margaret Morrison	51	67	.60	.65	.656	.644
Industries	55	70	.613	.606	.606	.634
Engineering	59	805	70	70	762	743

Part A represents familiar material, Part B the symbolic material, Part C the unfamiliar material, and Part D the suggestive material.

The relationship between success on the syllogism test and success in scholarship was measured by computing the coefficient of correlation for scores on the syllogism test and "points for quality" gained during the Freshman year. The Pearson Product-Moment formula was used throughout. The results are given in the following table.

<i>School</i>	<i>Number</i>	<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Total</i>
Margaret Morrison	47	416	.084	.37	.15	.406
Industries	49	226	.09	.32	.476	.41
Engineering	67	347	.286	.039	.192	.21

The scores on the syllogism test were also correlated with scores on the Thurstone Tests for Engineering Students in the case of 95 Sophomores of the School of Industries and the School of Engineering. The coefficients of correlation are given in the following table. The blank spaces in this table indicate that the scatter diagram showed no signs of relationship and no coefficient was computed.

<i>Thurstone Tests</i>	<i>Syllogism Test</i>				
	<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Total</i>
Arithmetic Test I	493	302	38	61	54
Algebra Test II		247	214	187	
Geometry Test III	35	25	25	38	32
Intelligence Test IV	36	51	40	44	53
Physics Test V	478	456	372	465	55
Technical Information Test VI	27		222	.362	

The table of mean scores for the different kinds of material on the syllogism test shows that Part A, the familiar material, is very much easier than the other material. Part D, the suggestive material, is more difficult than Part A, but not as difficult as the symbolic material or the unfamiliar material. Parts B and C are equal in difficulty. It may be noted that the engineering students show a decided superiority to the other students on all parts of the test. It might be expected that this superiority would show itself more in Part B, as the engineering students are supposed to have more familiarity with symbolic material than the students from the other schools. Their superiority in Part B, however, is no greater than their superiority in the other parts.

The table of correlation coefficients showing the relationship between success in the syllogism test and success in scholarship for the Freshmen year indicates a rather definite positive relationship between the total syllogism test and scholarship, especially in the Margaret Morrison School and the School of Industries. The group is such a highly selected one (college students who have succeeded in remaining in college well into the second year) that high correlations would not be expected between any test and scholarship. The coefficients for the different parts of the test and scholarship vary much. Part A has the highest coefficients and Part B the lowest.

The correlations between the syllogism tests and the Thurstone Tests for Engineers also vary considerably. A definite positive relationship is shown between the syllogism tests and Tests I, IV, and V—Arithmetic, Intelligence, and Physics. They show very little relationship between the syllogism tests and Tests II and VI—Algebra and Technical Information. With Test III, Geometry, the coefficients are fairly low, but rather consistent for the different parts of the syllogism test. The low correlation with the algebra test might be partly accounted for by the fact that the scores on this test piled up at the low end to such an extent as to make a high degree of correlation impossible. Of the different parts of the syllogism test, Part B shows the highest correlation with the intelligence test. Parts A and D show the highest correlation with the other tests.

There was no measure of the reliability of any of the test results and it was felt that the reliability would have to be determined before conclusions could be drawn from the results. In order to get a measure of the reliability of the scores, a second form of the syllogism test was devised, having items similar to those in the first form, but differing in slight logical detail, slightly different syllogistic forms being used, and of course differing in the concrete form and terms in which the syllogism was expressed. As it was felt that in the first form of the test there was no way to determine how much time was spent on the different material, the test was rearranged with the items of each kind of material in a separate booklet. It was thought that a wider range of ability would be found among high school students than among college students, and that a higher self-correlation would be found with such a group. The two forms of the test were therefore given to 76 pupils of the third and fourth years of high school. Twenty of these subjects were students in the Lincoln School, New York City, and fifty-six were students in a girls' high school in Brooklyn. The two forms of the test were given two weeks apart. Ten minutes were allowed for each part of the test, exclusive of the time allowed for reading instructions and studying the example given with the instructions.

The self-correlation showed that there was little reliability in the scores, especially the scores for Parts B and C. The

total test and Parts A and D show fairly high self-correlation. The results were as follows:

COEFFICIENTS OF CORRELATION BETWEEN THE TWO FORMS
OF SYLLOGISM TESTS

<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Total</i>
66	305	215	74	78

For the Lincoln School subjects taken alone the self-correlations are somewhat higher. The Method of Rank Difference was used in obtaining these coefficients.

COEFFICIENTS OF CORRELATION FOR LINCOLN
SCHOOL STUDENTS

<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Total</i>
70	64	42	79	872

An examination of the scores indicated that lack of reliability was to some extent due to the very large number of zero scores or scores near zero. When the subject's score showed almost as many items marked incorrectly as correctly, it seemed probable that the exact score was a matter of chance. With these high school subjects a large number of the scores were of this nature. In Part B, 40 scores out of 76 were below zero. The Lincoln School scores were somewhat higher and the reliability was higher.

The two forms of the test were then given to a class in psychology in the Extension Division of Columbia University. The results still showed a large proportion of zero and near zero scores and low self-correlations for the different parts, especially for Parts B and C. The method of Rank Difference was used to obtain these coefficients. The self-correlations were as follows:

<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>	<i>Parts A and D together</i>
68	28	56	73	85

It was clear that the test results would not be reliable unless the test was made easier or given to subjects more highly selected in regard to this particular ability. Since the Parts A and D taken together were much more reliable than when taken separately, it was thought that each part should be made longer; that is, should contain more items to be judged. If this were done, more time would have to be allowed for each part. These changes were made in the main experiment described in the next chapter.

CHAPTER II

MAIN EXPERIMENT

In this experiment the materials for the test and the selection of subjects were modified as suggested by the results of the preliminary experimentation. The test was lengthened so that each part consisted of sixty items instead of thirty-two and a twenty minute period was allowed for each part instead of the ten minute period. In a way the test was made easier in the new form in that three conclusions were given under each set of premises, and judging them is easier than if three different syllogisms had to be considered. The test was given to a more highly selected group as regards intelligence, the subjects being students in Columbia College and chiefly from the three upper classes.

1. METHODS

Subjects

The subjects for this experiment were eight-one undergraduates of Columbia College. Nine were Freshmen, twenty-three were Sophomores, fifteen were Juniors, and twenty-one were Seniors. The remainder had mixed classification or had failed to state their classification. The only principle of selection was interest in taking the test and, probably more important, desire to make the money offered for taking the test. In order to induce the subjects to do their best, the experimenter explained to them that the investigation would be valueless unless each one did his best.

That this group of undergraduates was representative of the whole undergraduate body as regards intelligence was indicated by a comparison of the distribution of scores on the Thorndike Intelligence Examination for this group and for the students entering in 1924 and 1925. The median score for the group in this experiment was two points higher than the median score for those entering. There was a slightly larger percentage of the experiment group with scores between 90 and 99, and a slightly smaller percentage of the experiment group getting scores between 70 and 90. On the whole the distributions are very similar.

Description of Test Material Used

The test material consisted of four parts: Parts A, B, C, and D, with two forms for each part. Each part consisted of twenty items numbered from one to twenty, each item consisting of two premises and three conclusions, a, b, and c, deduced from them. In some cases instead of two premises, just one proposition was used, with three conclusions drawn from this proposition. For instance, Part A, Number 14 is:

"All Freshmen take History I; therefore,

- a. all students taking History I are Freshmen;
- b. some students taking History I are Freshmen;
- c. some students taking History I are not Freshmen."

As each conclusion was to be marked by the subject, there were altogether sixty items to be judged in each part. Various forms of the syllogism were utilized and various common fallacies. Very unusual forms were avoided.

In Part A these syllogistic forms were embodied in familiar material, such as: "Some of the boats on the river are sailboats." For the most part, the conclusions, though concerning objects familiar to the subjects, were not actual facts within their experience; for the subjects would be liable in those cases to be influenced by their knowledge of the truth or falsity of the statement itself. For instance, "Mary's cats" is used instead of "cats" or "black cats."

Part B utilized the same syllogistic forms, but instead of familiar concrete terms, symbolic material, such as the letters, a and b, or x and y, were used. This might be considered more abstract and less familiar material than that used in Part A.

Part C consisted of exactly the same syllogistic forms, but the terms here were very unfamiliar, either scientific terms such as "echinoidea" or nonsense words invented to sound like scientific terms, such as "gyrofantastices."

Part D consisted also of the same syllogistic forms and the terms were familiar words such as those used in Part A, but the conclusions expressed facts within the experience of the subject and the truth or falsity of these statements was at variance with their validity as deduced from the given premises. A conclusion that could not logically be deduced from the given premises was an obviously true statement; or a conclusion that could be deduced logically from the given

premises was an obviously false statement. For instance, "Lincoln was an eminent man" is an obviously true statement; but in part D it was given as deduced from two negative premises. This might be called suggestive material, as the truth or falsity of the statements might tend to suggest that the conclusion was valid or otherwise, as deduced from the given premises. In some cases the fallacious nature of the conclusions was suggested merely by making the premises rather silly or nonsensical.

Not all the items of Part D were suggestive—for two reasons. If they were all of this nature, the subject might discover this after working over several items and might make a perfect score by marking all the true statements invalid and the false ones valid. The second and more important reason was in the impossibility of making three conclusions of this nature in the case of every syllogistic form used in the other parts of the test. For instance, Number 17 in Part D is as follows:

"Some sweet-scented flowers are red; all roses are sweet-scented; therefore,

- a. all roses are red;
- b. some roses are red;
- c. some roses are not red."

Here the premises are of such nature that no valid conclusion can be drawn from them in regard to the relation between the terms "roses" and "red" and therefore all three conclusions of these premises are invalid. The conclusions b and c are obviously true and so are intended to suggest validity. The conclusion a, however, is plainly a false statement and so whatever suggestiveness it might have would be necessarily toward marking it invalid. This conclusion, however, must be, of this form if it is to conform to this same item as it occurs in Parts A, B, and C.

Two-thirds of the items in Part D were designed to be suggestive in the sense described above. If the increase in difficulty of the item from Part A to Part D be taken as a measure of the suggestiveness of the item, a few of those designed to be suggestive proved not to be so; and a few of those not so designed proved to be suggestive. Obviously, the suggestiveness would vary much with the convictions of the subject. For example, the statement, "Some British are not Eng-

lish," was obviously true to the experimenter, but apparently had no such implications for the subjects.

The order of the items was varied in the different parts. For instance, Number 1 in Part A was Number 7 in Part B, Number 20 in Part C, and Number 17 in Part D.

The second form, Parts A*, B*, C*, and D*, was like the first, except that slightly different logical forms were used and, of course, the concrete terms were different. The same fallacies occur, though not with the same frequency.

The test material was mimeographed and put into booklet form, with Parts A, B, C, and D in separate booklets. The instructions were printed on the face sheet of each part. The instructions were identical for the different parts, but the example for Part B was in terms of letters and the example for Part C was in terms of long and unfamiliar words. The same example was used for A and D, and this was in familiar material. A sample face sheet is given herewith.

SYLLOGISM TEST

Write your name here

This is not an intelligence test but a test of a very special ability. In this test you are given certain statements followed by certain conclusions drawn from these. You are not concerned with the actual truth or falsity of these conclusions, but you are to determine whether, given these statements, you can draw the conclusions from them. In each case there are three conclusions given, marked a, b and c. You are to put a plus sign before every conclusion which you are sure follows necessarily from the given statements. You are to put a minus sign before every conclusion that does not necessarily follow from the given statements. If no conclusion can be drawn from the given statements, put a minus sign before each of the conclusions following.

The following example is correctly marked:

All tiksastopses are malpigienses; no malpigienses are tuscambia; therefore:

- +a. no tuscambia are tiksastopses.
- b. some tiksastopses are tuscambia.
- c. all tuscambia are tiksastopses.

Do not hurry. Accuracy counts more than speed.
Do not turn this page until the signal is given.

Procedure

The booklets for one part at a time were distributed to the subjects; time was given for reading the instructions and studying the example; and the experimenter read the instructions aloud. At a given signal the test was begun. Twenty minutes was given for each part of the test, exclusive of the time taken in reading the test instructions. The order of giv-

ing the parts varied somewhat. This is given in detail under "practice effect." The subjects took two parts in both forms in one evening and two parts in both forms in another evening. This meant a sitting of about an hour and a half each evening.

Scoring

The experimenter felt that the scoring should take into account accuracy and that this should be mentioned in the instructions, as otherwise many subjects would not take time to realize the difficulty of the task in hand, but would hasten through the test marking each item by a hurried impression and would perhaps be getting as many items incorrect as correct. The same individual, if not hurried, might give real thought to his responses and make a much more reliable score. The tests were scored then in two ways: first, the score equal to the number of items correctly marked; and second, the number of items correctly marked divided by the number of items attempted; or in others words, the per cent correct of those attempted. This latter method gave the higher reliability coefficients and therefore is judged the better method of scoring.

Reliability

The reliability of the syllogism tests used in this experiment is measured by computing the coefficients of correlation between the two forms of the test, thus obtaining a self-correlation for the test as a whole and for each part separately. The Pearson Product-Moment formula was used. The following table gives the coefficients of correlation for each part of the test, and for the total, and for the two methods of scoring.

COEFFICIENTS FOR SELF-CORRELATION OF SYLLOGISM TESTS

	Score $\frac{R}{R+W}$		Score No. R		
	<i>r</i>	<i>P.E.</i>	<i>r</i>	<i>P.E.</i>	<i>Number</i>
Total	.965	005	.915	.013	79
Part A	.90	014	.845	.022	80
Part B	.883	.018	.785	.029	80
Part C	.855	.021	.85	.021	80
Part D	.841	022	80	.027	80

The distribution of scores in the two forms is very similar as may be seen from the means of the scores for each part and

the standard deviations of the distributions. These are given herewith. The second form is slightly more difficult than the first form.

	<i>First Form</i>		<i>Second Form</i>	
	<i>Mean Score</i>	<i>S.D.</i>	<i>Mean Score</i>	<i>S.D.</i>
Total	79 35	13 7	77 3	13.7
Part A	85 27	13 6	83 85	13.6
Part B	77 68	15 7	73 56	14.1
Part C	76 37	15 2	73.94	14.
Part D	80 75	16 4	78 37	16 5

The coefficients of correlation are probably higher by a slight amount than the actual relationship warrants, due to the somewhat bi-modal distribution of the scores. Instead of massing of measures near the mean, there is a thinning out of measures there. This occurs in both forms of the test and to some extent in all parts. This will be discussed in more detail later.

In spite of this irregularity of distribution of scores, there is evidently a close relationship between success in the first form of the test and success in the second form. McCall⁸ states that the best intelligence tests have a self-correlation of form .90 to .95 and that most standard tests have a reliability of about .80. According to this statement, this test as a whole has a reliability that is high for a group of college undergraduates. The separate parts also are fairly reliable, the familiar material being the most reliable and the suggestive the least. This last is to be expected, as there could be no measure of the amount of suggestiveness of the different items. This would vary from individual to individual with each item. Also, in some cases, the subject might suddenly become on his guard against suggestiveness, and thereafter his reactions would be very different. This might occur anywhere within the test and would doubtless affect the reliability.

With the above standard in view, these tests are sufficiently reliable for the purposes of this experiment. This does not mean that these tests would be reliable when given to any group taken from the population at large. These subjects are a highly selected group as regards intelligence. On this test they yield a range of scores from .45 correct of the items attempted to a perfect score of 1.00. It is highly probable that with a less intelligent group the zero scores would be

sufficiently numerous to make a high coefficient of reliability impossible. Indeed, the evidence from the preliminary work is in favor of the test's being less reliable with a less intelligent group. This is especially true of Parts B and C, as they are the most difficult. Part A is rather too easy for this group of subjects, the high scores being very numerous, so that this Part A might be expected to prove reliable with a group of high school students.

Effect of Training on Distribution of Scores

As stated above, the curve of distribution of the scores in the syllogism test is very irregular and has a somewhat bi-modal form. This occurs in both forms of the test and to some extent in all parts. There is nothing in the results to suggest a reason for this. Possibly the members of the higher group have an analytical approach and some sort of method while the members of the lower group depend upon general impression or feeling for a decision.

One might be disposed to consider that this bi-modal distribution was due to difference in training, the higher group being those that had received training in this sort of reasoning; the lower group being those that had had no training. In order to get some information on this point, statements as to training were obtained from 44 of the subjects. They were asked to check answers in the following questions:

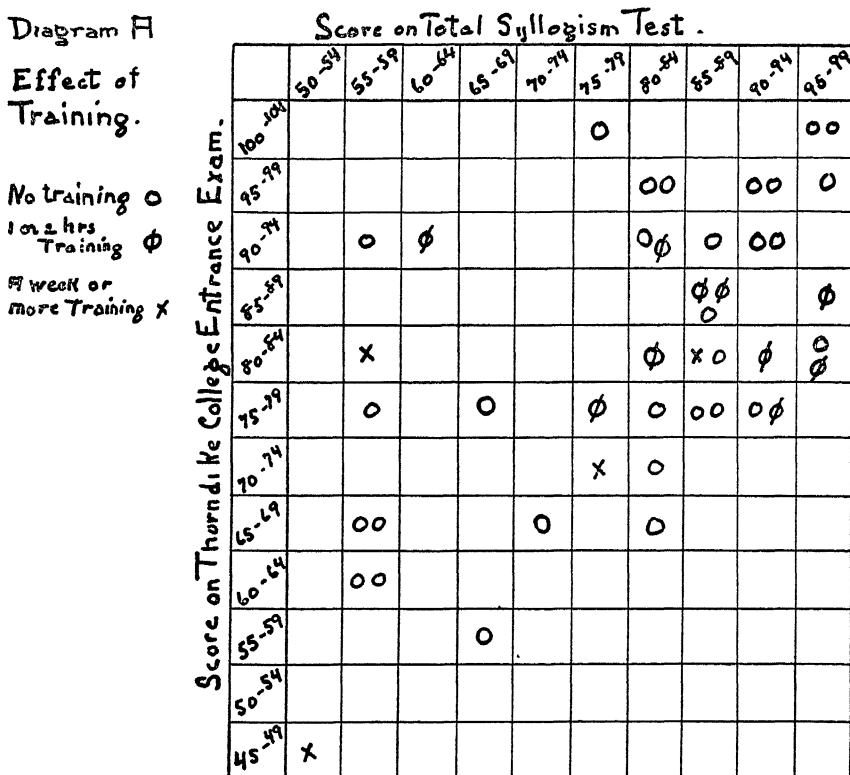
Before taking this test in syllogistic reasoning

1. Had you had a course in Logic? Yes No
2. Had you had as much as a week's training in this sort of thing in a psychology course? Yes No
3. In an English course? Yes No
4. Had you had one or two lessons devoted to this sort of thing in a psychology course? Yes No
5. In an English course? Yes No
6. Did some one give you a little training just before you came to take the test? Yes No
7. Have you had any other training not included in the above?
 Yes No
 If so, what?

Of the 44 cases from whom statements were obtained, only 14 admitted any training and only four of these admitted as much as a week's training. An hour's training in some Eng-

lish class was that most frequently mentioned. Thirty subjects said that they had no training whatever.

The scores of those admitting training are scattered throughout the distribution of the scores for the whole group. Two of the four who said they had as much as a week's training are at the lowest end of the distribution. Of the twenty-three cases scoring below 85 per cent on the syllogism test, seven admitted training; of the twenty-one cases scoring above 85 per cent, seven admitted training.



On the accompanying scatter diagram (A) the individuals having different amounts of training are represented by different symbols and are placed according to their scores on the syllogism test and their scores on the intelligence test. If training had affected the scores to a great extent, we would expect to see a large number of those having training to the right of the main trend of correlation, and a large number of

those without training to the left. This does not seem to be the case to any significant extent.

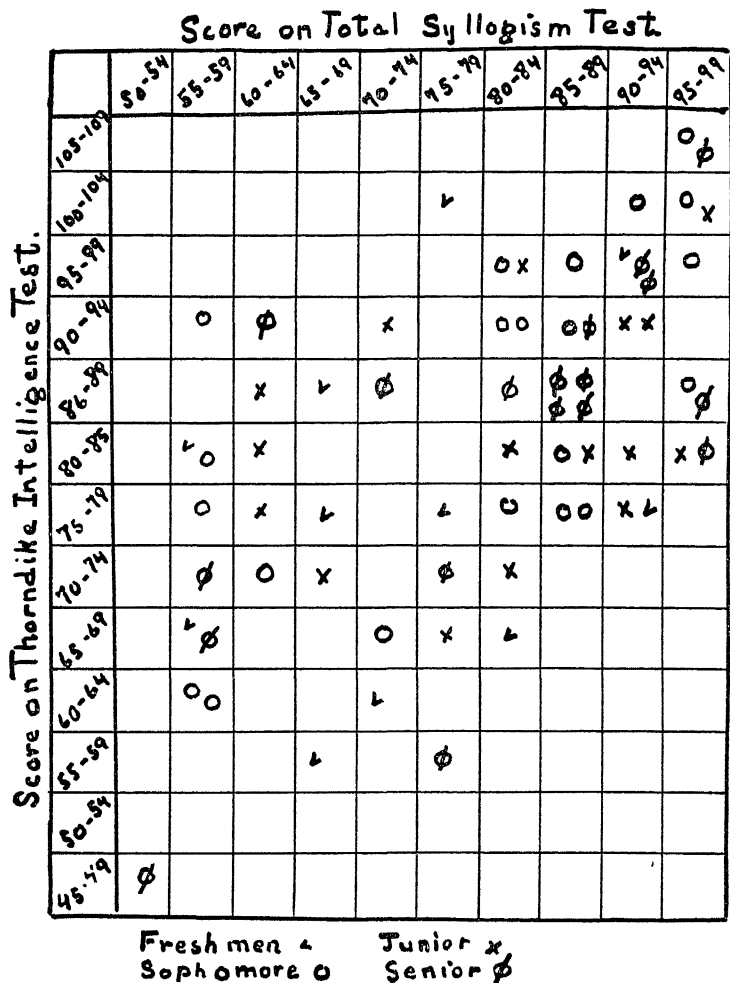
It seems probable to the experimenter that the subjects have all had some training in English courses, both in high school and college, but on a large number of them it made little impression. One subject said that the only training, if it could be called such, that he had had, was the statement of a high school teacher that the converse of a true proposition was not always true. One can imagine how many of the class in hearing that statement remembered it several years later. This subject must have been interested in that sort of thing and had profited greatly by having his attention called to that principle. The experimenter would judge that the subjects' training has been even more uniform than their statements would lead one to believe. As far as these results go they indicate that the irregular distribution of the scores is not due to differences of training; that we are not dealing with two groups—one trained and the other not.

Another bit of evidence bearing on this subject of training is the comparison of the mean scores for the syllogism test for the subjects classified as Freshmen, Sophomores, Juniors, and Seniors, with their mean scores on the intelligence test. We might expect that the college curriculum might give training in careful thinking and that the lower group on the syllogism test might be under classmen, and the upper group upper classmen. The accompanying table gives these data.

Class	1928	1927	1926	1925
Number	11	21	16	18
Mean Score on Syllogism Test (Total)729	.80	.815	.80
Mean Total Score on Intelligence Test776	.851	.836	.832

The mean score for the Freshmen in the syllogism test is decidedly lower than the mean scores for the other groups. The mean scores for the other groups are practically the same. When we consider the mean intelligence test scores for these groups we see that the Freshmen are a poorer group than the others in intelligence. It might be this rather than a lack of training that causes the low score on the syllogism test. The group of Sophomores have a higher mean score in the intelligence test than either the Junior or Senior group and yet their mean score in the syllogism test is about the same. This might be taken to mean that the Juniors and Seniors had had some training not had by the others that had helped them on the

syllogism test. This, however, is not sufficient to explain the distribution.



Diagram

(B) Effect of College Classification on Place in Test.

The accompanying scatter diagram (B) shows the individuals belonging to the different classes as their scores in the two tests place them. If the scores on the syllogism test were much affected by the training given by college work, we would expect Seniors and Juniors to predominate in the lower right side of the diagram, and the Freshman and Sophomores in the

upper left. There is a very slight tendency shown toward this, but it is not sufficiently marked to be in any degree conclusive and is certainly not sufficient to account for the bi-modal distribution.

Practice Effect

The test was so long that there was ample room for practice to take effect. In order to obviate this difficulty to some extent and to get an inkling of how great the effect would be, the order of the parts was somewhat varied. The numbers are too small to give any definite conclusions, but there is some evidence that the gain from practice is about 5 points on the scale of 55 points, about 9 or 10 per cent.

One group had Part D first and Part A last. The other groups had Part A first and Part D last. The mean score on Part D when taken first is 5 points lower than the mean score when taken last. The same is true of Part A. This suggests that if the effect of practice were ruled out, the average score of Part A would be higher and the average score of Part D would be lower, thus increasing the difference between Part A and Part D.

Parts B and C were never placed first or last, but in two groups B was placed second and in two groups third. The evidence here is not so clear as with A and D, but shows a rise of 2 or 3 points due to practice.

Thorndike¹² reports a median gain of 12.3 per cent from first to second trials of an intelligence test given to college students. Dunlap and Snyder⁴ report a median gain from first to second trials of the Army Alpha given to 44 college students as 15 points in a range of 107 points, about 14 per cent. These are not exactly parallel cases to the investigation at hand, but serve to show that the gain in practice reported here is not excessive

2. RESULTS

Difference in Difficulty Caused by Change of Material

a. The main problem of this investigation was to determine the effect on ability to do formal syllogistic reasoning, of changing the material in which the syllogistic forms were expressed. In order to see whether the relative standing of individuals remained the same on different parts of the test, correlation diagrams* were constructed and coefficients com-

* Representative correlation diagrams may be found at the end of this article.

puted as given in the following table. The Pearson Product-Moment formula was used throughout.

In order to see what the relationship between different parts of the test would be if the factor of intelligence were ruled out, the scores on the Thorndike College Entrance Examination were used as a measure of intelligence, and the method of partial correlations was used. The coefficients of correlation, obtained when the intelligence test factor is rendered constant, are given in the last column of the following table.

b. In order to determine the difficulty of the different parts of the test, the arithmetical mean was computed for each part of the test in its two forms and the distribution of scores was tabulated. These are given in the following table. The means given in this table differ slightly from the means given above under the head of reliability. This difference is due to the fact that the means here given were computed from the sum of each individual score and those given in the former table are computed from the measures as grouped in class intervals.

CORRELATION BETWEEN DIFFERENT PARTS OF THE TEST

	<i>r</i>		<i>Number</i>
Parts A and B70	.038	80
Parts A and C76	.032	80
Parts A and D685	.039	80
Parts B and C805	.026	80
Parts B and D75	.033	80
Parts C and D73	.035	80

When the cases having no score on the intelligence test were omitted, the results were as follows:

	<i>r</i>	<i>Intel. Constant.</i>	<i>Number</i>
Parts A and B68	.548	69
Parts A and C76	.630	69
Parts A and D685	.583	69
Parts B and C814	.740	69
Parts B and D754	.653	69
Parts C and D680	.680	69

The following formula for calculating partial correlations in the case of three variables was used.

$$r_{12, 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

MEAN SCORES FOR EACH PART OF TEST

Score = % Right		Score = Number Right	
Part A	.85 \pm 1.01	Part A	50.5 \pm .7
Part A*	.837 \pm 1.02	Part A*	49.8 \pm .66
Part B	.78 \pm 1.17	Part B	44.8 \pm .76
Part B*	.738 \pm 1.03	Part B*	41.4 \pm .69
Part C	.754 \pm 1.14	Part C	41.3 \pm .85
Part C*	.738 \pm 1.05	Part C*	39.2 \pm .65
Part D	.795 \pm 1.23	Part D	46.6 \pm .76
Part D*	.776 \pm 1.24	Part D*	44.9 \pm .74

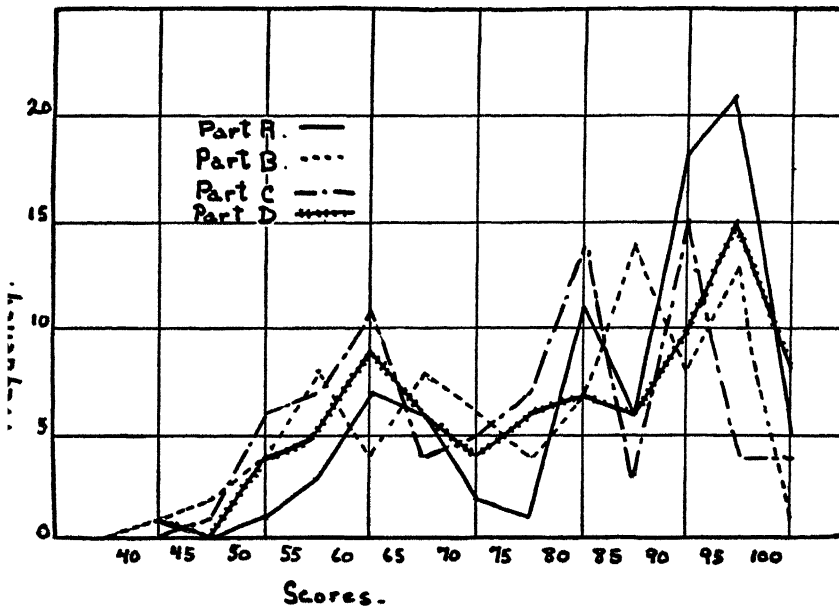
C. DISTRIBUTION OF SCORES ON EACH PART OF THE SYLLOGISM TEST

Score = % right of those attempted

<i>Class Intervals</i>	<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>
100	5	1	4	8
95-99	21	13	4	15
90-94	18	8	15	10
85-89	6	14	3	6
80-84	11	7	14	7
75-79	1	4	7	6
70-74	2	6	5	4
65-69	6	8	4	6
60-64	7	4	11	9
55-59	3	8	7	5
50-54	1	4	6	4
45-49	0	2	1	0
40-44	0	1	0	1

Score = number right

60	5	1	2	8
55-59	33	15	11	15
50-54	17	21	10	15
45-49	6	4	12	7
40-44	6	12	7	9
35-39	8	12	15	17
30-34	4	8	10	7
25-29	1	6	8	3
20-24	1	1	5	0
15-19	0	0	1	0

(C) Distribution of Scores. $S = \frac{R}{R+W}$ 

3. CONCLUSIONS

There is a high degree of relationship between ability to reason with familiar material and ability to reason with more abstract material, such as the symbolic material in Part B, the unfamiliar in Part C, and the suggestive or deliberately misleading material of Part D. Since the coefficients of correlation between the different parts of the test are not as high as the coefficients between two forms of the same part, it would seem that changing the material does to some extent change the position of some of the individuals in regard to their ability to do this sort of reasoning. That is, some individuals do relatively better with more abstract material than with more familiar and concrete material, and others do relatively better with familiar material.

Some degree of this correlation measured by the coefficients given is undoubtedly due to differences of intelligence among the subjects. We find, however, that when this is allowed for, and by the method of partial correlations the degree of relationship is determined that would exist if we had all the sub-

jects of the same degree of intelligence, there still exists a fairly high correlation expressed by the coefficients .548, etc.

The ability to reason with symbolic material is more closely related to ability to reason with unfamiliar material than to ability to reason with familiar, more concrete material. The ability to reason with familiar material when it is suggestive is more closely related to ability to reason with symbolic material than with ability to reason with familiar material that is not suggestive. It is probable that the processes involved in dealing with the material of Parts B, C, and D are more alike than the processes involved in Part A are like any one of them. These three parts all involve more abstraction, more ability to resist distraction than Part A does. They are all less familiar than Part A and involve less habitual reactions.

The average scores for the different parts of the test and the distribution of scores show clearly that in general it is much easier to reason with familiar material than with the unfamiliar and symbolic. Part C (the long words) was even more difficult than Part B (the symbolic material). Even when the familiar material was made suggestive, it was easier than the symbolic and the unfamiliar material.

These differences are more clearly seen from the table showing the distribution of scores than from the table of average scores. It is clear that Part A has very many more high scores and fewer low scores than the other parts. It is clear from the distribution that Part A is not hard enough to measure the ability of the subjects at the higher levels. Part A, however, could not be made more difficult without increasing alarmingly the number of zero scores on the other parts.

If we consider the position of individuals in regard to the different parts of the test, we find that 12 subjects found Part A more difficult than Part B and 52 subjects found Part B more difficult. Of the 12 who found A more difficult, only 3 had scores for the two parts differing by more than 10 points. Of the 52 who found B more difficult, 23 had scores for the two parts differing by more than 10 points, and 17 had scores for the 2 parts differing by as much as 20 points, 1 had a difference of 35 points and 1 a difference of 45 points. In other words, a few individuals who did well on Part A went all to pieces on Part B. Almost half of those finding B harder found it very much harder. Those that found A more difficult, did not find it much more difficult. When the dif-

ferences are not great, chance probably is somewhat responsible for them.

In comparing Parts A and C in the same way, we see that 9 cases find Part A more difficult and 59 find C more difficult. Where the difference is in favor of Part A, it is very small, probably not significant; but where C is the more difficult, in many cases the differences are very large. In two cases the difference is as much as 35 points, in 7 cases the difference is as much as 25 points, in 29 cases the difference is more than 10 points. This is the same sort of situation that we find in the case of Parts A and B.

When we compare individuals in the same way with reference to their scores on Parts B and C, we find a very different situation. Here nearly as many find B the more difficult as find C the more difficult. Twenty-eight cases make a higher score on Part C than on Part B, and 33 cases make a higher score on Part B than on Part C.

Parts A and D compared in the same way show a rather similar relationship to that of Parts A and B. Seventeen cases find A more difficult, but the differences are small, only two cases having a difference of as much as 20 points. Thirty-eight cases find Part D more difficult than Part A, and in many cases the differences are large. In 8 cases the difference in scores is as much as 25 points. That is, several subjects who did well in Part A did very poorly in Part D.

In whatever way we examine the data, then, it is clear that in general it is easier to deal with familiar material than with any of the other material. The ability to do formal reasoning is much affected by changing the material that is reasoned about from familiar to symbolic or unfamiliar, or by introducing suggestion into the familiar material. It is probable that habit is responsible for this. The subjects are much more accustomed to dealing with material of the Part A type than of any of the other types. Part C is more difficult than Part B, because the subjects have some familiarity with these symbols in their mathematics, etc., but have very little familiarity with such words as *Tikthostopses* or *Siscumbaba*. Later in this discussion, when we consider certain particular items which are more difficult in Part A than in Part B, this point comes out clearly. Many of the subjects probably are guided in marking the items of Part A by a feeling of rightness or wrongness according as it fits in with their habits of thought

or not. They can often tell whether the item is fallacious when they are unable to explain why. When these subjects contend with the symbolic material or the big words, they no longer have this feeling to guide them and, if they are to be successful in the test, they have to hit on some principles to help them. When they attempt to work on Part D, these vague feelings of rightness and wrongness are apt to be on the side of the truth or falsity of the statement, rather than on the side of logical or illogical deductions from the premises.

On the other hand, certain individuals probably perceive the familiarity of the form even when the content is unfamiliar, and these individuals show relatively little variation in ability from one part of the test to another. These individuals are capable of analysing the problem more completely than those who are much affected by the unfamiliarity of the material, and see elements of the situation to which the others are blind.

4. CORRELATION BETWEEN SYLLOGISM TEST AND THORNDIKE COLLEGE ENTRANCE TEST

In order to determine in some measure the relationship between intelligence and ability to do formal syllogistic reasoning, and especially to see whether this relationship is affected by changing the material of the syllogistic forms, the coefficients of correlation were computed between the scores on the Thorndike College Entrance Test and scores on the Syllogism Test, both the total test and each part. In regard to the Thorndike College Entrance Tests, Pintner⁹ says, "The test is very much more comprehensive and much more difficult than the Army Alpha Test. It requires two hours and fifty minutes actual working time and includes some educational tests of a type suitable for high school graduates. A correlation of .65 between the work of the entire Freshman year and the scores on the mental test is reported by Thorndike." A full description of this test is given by Ben D. Wood¹³ in his "Measurement in Higher Education." The subjects in this experiment had taken this test when seeking admission to Columbia College and it was the scores obtained at that time that were used in the present study. Scores on the Thorndike Test could be obtained for only 69 of the subjects taking the syllogism test. The others had entered with advanced standing or for some other reason had not taken the intelligence

test. The results on this part of the study, then, are limited to 69 subjects. The correlations were obtained with the two kinds of scores for the syllogism tests, in one case the score being per cent right of those attempted and in the other case, the number right. The coefficients of correlation were obtained by means of the Pearson Product-Moment Formula, and are given in the following table:

CORRELATION WITH THORNDIKE COLLEGE ENTRANCE TEST

	Sc. $\frac{R}{R+W}$	Sc. No. R.
Total Syllogism Test and Thorndike	.578 \pm .054	.505 \pm .060
Part A and Thorndike	.498 \pm .061	.426 \pm .066
Part B and Thorndike	.596 \pm .053	.613 \pm .050
Part C and Thorndike	.505 \pm .060	.39 \pm .069
Part D and Thorndike	.485 \pm .062	.486 \pm .062

Conclusions

These coefficients show a definite positive relationship between ability to do formal syllogistic reasoning as measured by these syllogism tests and general intelligence as measured by the Thorndike College Entrance Examination. According to Rugg,¹⁰ correlation is "markedly present" or "marked" when r ranges from .35 or .40 to .50 or .60; is "high" when it is above .60 or .70. When one takes into consideration that the group represented here is a very highly selected one as regards intelligence, and that each part of the syllogism test covered a period of only twenty minutes, these correlations seem fairly high. Kitson⁷ obtains coefficients ranging from .18 to .60 between standings in various tests such as opposites, hard directions and memory for logical material and standings in the net score of all the tests together. Arlitt¹ reports a correlation of only .23 between intelligence quotients obtained from the Stanford Binet Scale and scores on the Thurstone Intelligence Test given to a group of college Freshmen.

This degree of relationship means, however, whether we call it high or low, that there are many individuals having very different standings in the two tests and that we cannot predict with any sureness the place of an individual in one test if we know his position in the other. The correlation diagrams show several subjects that stand much higher in the syllogism test than in the intelligence test. These are mostly near the median score in the intelligence test and in the highest quarter of the syllogism test. On the other hand, much more con-

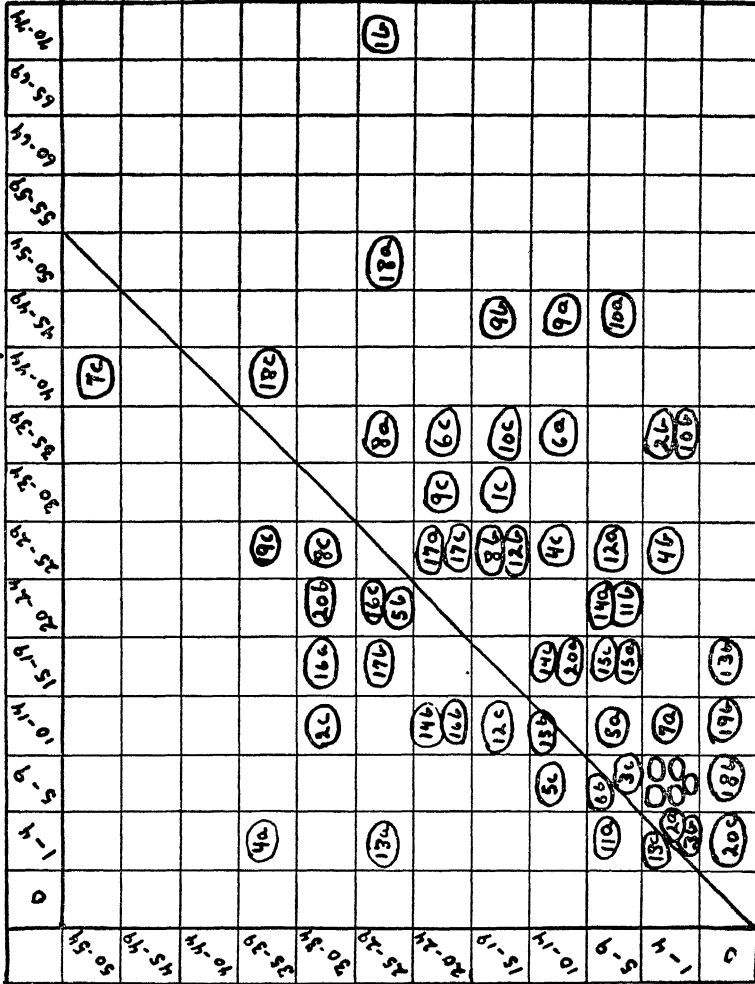
spicuously out of line with the general trend of the correlation table is a number of cases that seem to form almost a separate group. These are mostly just above the median of the intelligence scores and very low in the syllogism test. These seem to be individuals of good general intelligence who are able to succeed but poorly in formal syllogistic reasoning. One individual who on the intelligence test has only two individuals making a better score, on the syllogism test has 39 individuals making a better score. This group stands out conspicuously on all parts of the test except Part B. If the diagram (G) is examined it will be seen that in Part B the scores of these individuals for the most part were not actually higher than in the other parts, but only relatively so.

The correlation between scores on Part B and scores on the intelligence test is much higher than that between any other part and the intelligence test. This is not surprising, for one would expect ability to handle symbols to play an important part in general intelligence at the upper levels. It is difficult, however, to see why Part C, which correlates so closely with Part B, should have so much lower a correlation with the intelligence test than Part B has.

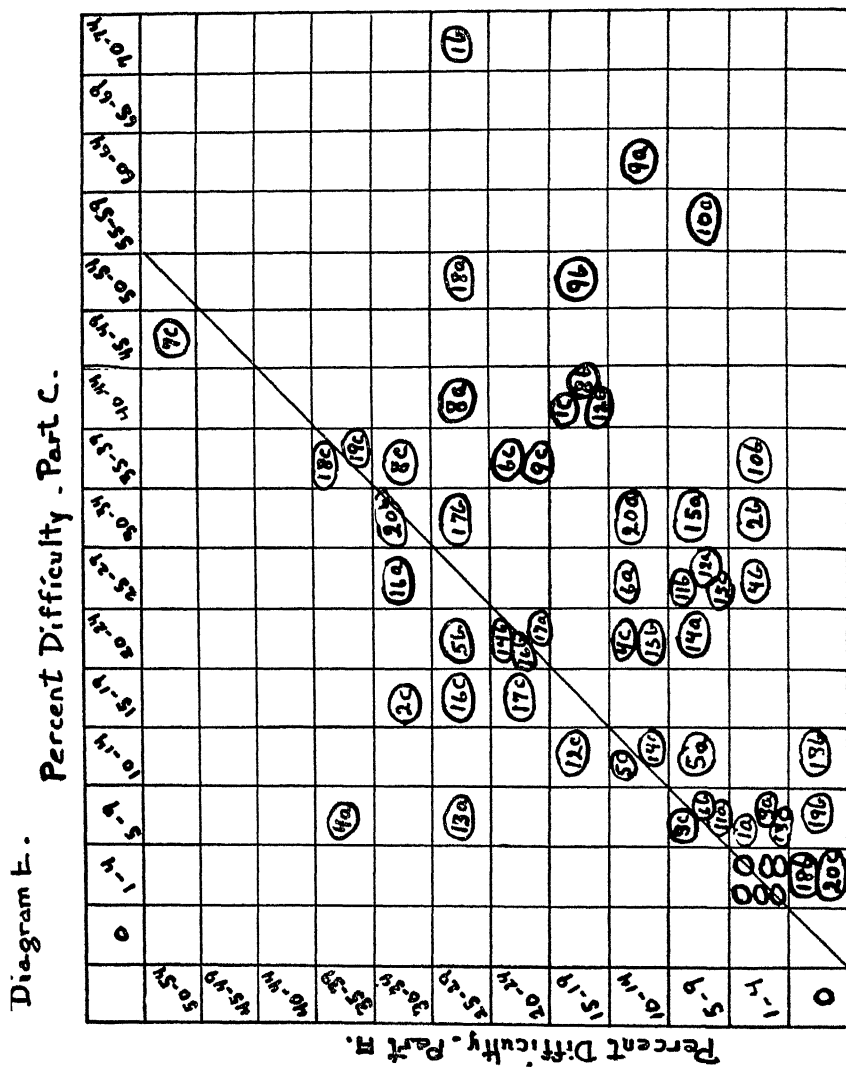
Perhaps Part A would show a higher correlation with the intelligence test if it were not for the piling up of scores at the higher end of the scale. As has been stated before, Part A is too easy to distribute the scores properly at the upper end.

Diagram D.

Part B - Percent Difficulty.



Part A - Percent Difficulty.



CHAPTER III

EFFECT OF CHANGING THE MATERIAL ON THE DIFFICULTY OF EACH ITEM AND EACH KIND OF ITEM

1. DISCUSSION

As the different items involved various sorts of fallacies, it is interesting to see which are most difficult for the college student and how their difficulty varies in the different parts of the test. For this purpose the per cent marking the item incorrectly of the number of subjects attempting it was calculated, and this was called its per cent of difficulty. The items in each part of the test were then ranked according to this difficulty; the most difficult, that is, the one having the largest per cent difficulty was given a rank of 1. It is clear, then, that ranks below 30 represent the more difficult items and ranks above 30 the easier ones. These ranks are given in the accompanying table. The items are grouped in this table according to the sort of logical fallacy involved. The items involving no fallacy are grouped together under the head of "Valid." Since many items involve more than one fallacy, the same item appears in several groups.

Before considering the kind of fallacy involved, the items were plotted on a scatter diagram (D) placing them as to their difficulty in Parts A and B. A similar diagram (E) was constructed for Parts A and C. The figures and diagrams give the following information as to the change in difficulty effected by change of material.

Discussion of Results

Forty-three items are easier on Part A than on Part B, 2 are neither harder nor easier, and 15 are harder on Part A. of these 15, 5 are very nearly the same in both, leaving 10 that are decidedly easier on Part B. Of the 43 items easier on A, one increases in difficulty 45 points, and 9 increase as much as 20 points.

When the items are examined as to their difficulty in Parts A and C, the same results are found in slightly exaggerated form. Forty-five items are harder on C than A; 13 items are harder on A. Many change very little but only 4 stand out as

harder on A and 25 stand out as harder on C. On 3 items the difficulty increased 50 points on the per cent scale. On 11 items the increase was as much as 25 points.

No similar diagram was made for the items of Parts A and D, as the suggestiveness of the items was such a varying quantity. Of the items intended to be suggestive 7 failed to be, if suggestiveness is measured by increased difficulty in Part D over Part A. The following proved to be the most suggestive.

6 b. All Anglo-Saxons are English; all British are Anglo-Saxons; therefore, all English are British.

Only 1 per cent marked this item incorrectly—in the form in which it appeared—in A, and 32 per cent marked it incorrectly in D.

9 a. No people interested in modern drama have failed to read this book; no people who have failed to read this book are actors; therefore, all actors are interested in modern drama.

The difficulty of this item was increased 44 points.

11 a. All monkeys have tails; human beings are not monkeys; therefore, human beings have no tails.

The difficulty of this item was increased 17 points, 13 per cent marking it incorrectly in A and 30 per cent marking it incorrectly in D.

14 a and 14 c. No oranges are apples; no lemons are oranges; therefore,

a. no lemons are apples;

c. no apples are lemons.

In Part A, 6 per cent marked a incorrectly and 6 per cent marked c incorrectly. In Part D, 31 per cent marked a incorrectly and 31 per cent marked c incorrectly.

Items 12 a, 12 b, and 12 c were intended to be suggestive but apparently were not so, as they were about equally difficult in Parts A and D. The experimenter considered the conclusions so obviously true that she expected the subjects to be influenced by this to mark them correct.

12. Some books are not worth reading; some books are not allowed in circulation; therefore,

a. some books allowed in circulation are not worth reading;

b. some books worth reading are allowed in circulation;

c. some books worth reading are not allowed in circulation.

When we examine the tables of items grouped according to the kinds of fallacies involved, we get some interesting light

on the items that stand out on the above diagram as easier on Part A than on Parts B and C. Three of the conspicuous ones, 4 a, 13 a, and 17 b, are of the same nature, a positive conclusion drawn from two negative premises. 13 a is typical.

13. No good physician advertises his cures; Dr. J. does not advertise his cures; therefore, Dr. J. is a good physician.

Twenty-seven per cent of the subjects marked this item incorrectly, but when it was reduced to symbolic form, only 3.8 per cent failed to see the fallacy. Item 4 a was marked incorrectly by 36 per cent of subjects as it appeared in Part A, and by only 3.8 per cent as it appeared on Part B. Three other items, 2 c, 19 c, and 14 b, while they were not quite so conspicuously harder in Part A, but somewhat so, were all negative conclusions from positive premises. For instance, 2 c is: Some of John's rare books are on this shelf; therefore, some of the books on this shelf are not John's rare books. Three others easier on A, 16 a, b, and c, are negative conclusions from positive premises. One of these, 16 b, is valid.

It is probable that the fallacies involved in these items are very usual ones in daily life, so that when they are expressed in familiar terms, the subjects feel that they are valid, and do not stop to examine them. When the same thing confronts them in symbolic or unfamiliar material, they are forced to consider the relationships involved and so perceive the fallacy.

From this consideration of the changes in the several items from part to part of the test, it is clear that changing the material away from the familiar adds much to the difficulty except in the cases just discussed, where bad habits of reasoning interfere.

2. TABLES

a. Explanation of the Table of Rank Difficulty and Diagnostic Value of Items

The separate items of the test are classified according to the kind of fallacy involved. Many items appear more than once in the table because more than one fallacy is involved. Across the page the same item appears as it occurs in the different parts of the test, familiar material, symbolic, etc. The name of the item varies from part to part of the test because the order of the items was varied in the different parts. With the name of each item is given its rank difficulty in that part

of the test, the diagnostic value of that item for the test and its diagnostic value for intelligence. Under the head of remarks are put in abbreviated form the names of other fallacies involved or any condition that might affect the difficulty. By reading across the page one sees how the item varies in all these respects as the material changes. By reading the vertical columns one sees the difficulty, etc., of items in each kind of fallacy.

The first of these tables includes the items of the first form, the second the items of the second form. The results from these two tables are averaged and these averages are given in a table following.

RANK DIFFICULTY AND DIAGNOSTIC VALUES OF TESTS—FIRST FORM

Part A				Part B				Part C				Part D				Remarks
Name	Rank cultry	Diag Val. for Test	Diag Val. for Intel	Name	Rank cultry	Diag Val. for Test	Diag Val. for Intel	Name	Rank cultry	Diag Val. for Test	Diag Val. for Intel	Name	Rank cultry	Diag Val. for Test	Diag Val. for Intel	
18c	4	2 00	1 24	17c	6	7 00	2 38	17c	13	3 50	1 08	18c	20½	2 13	1 48	Neg from Positive
17a	21	2 80	1 40	15a	23	2 46	1 82	1a	18	3 20	1 09	12a	27½	1 93	1 28	2 partics
17b	15	3 50	1 66	15b	34	1 54	1 24	1b	33	2 04	1 58	12b	16	2 76	1 36	2 partics
Fallacy of Illicit Minor Term																
2b	54½	1 11	1 09	2b	9½	2 76	1 79	13b	21	2 71	1 77	8b	11	3 70	1 68	
6c	18	1 85	1 23	6c	11½	3 93	1 79	9c	15	3 70	1 31	7c	22½	2 34	1 31	
10a	44	1 31	1 06	9a	4	3 65	1 79	8a	3	3 65	1 31	20a	26	2 14	1 76	
12b	25	1 55	1 28	12b	21½	3 34	1 92	7b	8	3 20	1 79					
Fallacy of Illicit Major Term																
6b	38½	1 46	1 16	6b	47½	1 32	1 03	9b	46	1 16	1 03	7b	46	1 22	1 23	Neg from Positive
9a	30½	1 65	97	14a	5	3 05	1 36	2a	2	2 80	1 23	11a	14	3 00	1 39	
9c	19½	1 96	86	14c	14	1 11	73	2c	12	1 10	72	11c	29½	1 50	97	
10c	23	2 30	1 45	9c	13	92	1 03									
12c	27	2 00	94	12c	41	1 16	1 16	7c	40½	1 05	97					Neg from Positive
Fallacy of False Conversion																
14a	38½	1 46	1 06	8a	24½	2 76	1 48	15a	33	1 88	1 18	5a	22½	2 34	1 23	
19c	2	5 00	1 60	19c	18	2 74	2 00	16c	16	2 51	1 28	15c	15	2 62	1 46	Obverted
16a	6	3 80	1 70	16a	33	2 62	2 02	11a	25	2 15	1 24	3a	27½	1 55	1 41	Obverted
19a	47	1 25	1 17	19a	49	1 40	1 50	16a	60	.97	97	15a	56	1 00	1 06	Universal from P

RANK DIFFICULTY AND DIAGNOSTIC VALUES OF TESTS—FIRST FORM

Part A				Part B				Part C				Part D				Remarks
Name	Rank Diffic- ulty	Diag. Val	Diag for Syl. for Intel	Name	Rank Diffic- ulty	Diag. Val	Diag for Syl. for Intel	Name	Rank Diffic- ulty	Diag for Syl. for Test Intel	Diag Val	Name	Rank Diffic- ulty	Diag for Syl. for Test Intel		
Conclusions Drawn from Two Negative Premises																
4a	3	3 35	2 04	4a	55½	1 19	1 16	3a	48	1 32	1 20	2a	1	4 70	1 58	Pos f Neg
4b	29½	1 17	92	4b	18½	2 96	1 58	3b	26½	2 76	1 28	2b	50½	1 10	.97	
4c	30½	1 17	97	4c	24½	3 16	1 68	3c	33	1 92	1 54	2c	42	1 58	1 23	
11a	41½	1 38	1 26	13a	57½	1 14	1 00	4a	46	1 42	1 09	9a	3	4 70	1 31	Pos f Neg
11b	41½	1 24	97	13b	28	2 20	2 00	4b	23	2 15	1 54	9b	57½	1 05	1 03	
11c	29½	1 17	1 12	13c	45	1 58	1 21	4c	57	1 02	94	9c	44	1 35	1 03	
13a	11	3 56	1 64	11a	54	1 11	1 16	6a	40½	1 58	1 16	6a	5	5 15	1 91	Pos f Neg.
13b	58½	1 00	1 00	11b	30	2 08	1 39	6b	40½	1 58	1 16	6b	59½	1 00	97	
13c	52	1 18	1 10	11c	57½	1 14	1 09	6c	46	1 42	97	6c	59½	1 00	97	
15a	43	1 40	1 10	18a	31	2 58	2 05	1a	18	3 20	1 09	14a	12	2 66	1 32	
15b	29	2 00	1 10	18b	38	1 68	1 28	1b	33	2 04	1 58	14b	25	1 33	1 09	Some not
15c	40	1 32	97	18c	36	2 20	2 00	1c	24	2 15	1 03	14c	13	2 93	1 44	
17a	21	2 80	1 40	15a	23	2 46	1 82	18a	30½	1 16	89	12a	27½	1 93	1 28	2 Part.
17b	15	3 50	1 66	15b	34	1 54	1 24	18b	20	2 63	1 45	12b	16	2 76	1 36	Prem's
17c	17	3 18	1 37	15c	16	3 30	1 87	18c	36	1 42	75	12c	36	1 93	1 13	2 Part
Positive Conclusions from Negative Premises																
4a	3	3 35	2 04	4a	55½	1 19	1 16	3a	48	1 32	1 20	2a	1	4 70	1 58	Both Prem's
7b	54½	1 11	1 03	1b	51	1 19	1 03	14b	54½	1 06	88	1b	57½	1 05	1 03	Negative
11a	41½	1 38	1 26	13a	57½	1 14	1 00	4a	46	1 42	1 09	9a	3	4 70	1 31	Both Prem's
13a	11	3 56	1 64	11a	54	1 11	1 16	6a	40½	1 58	1 16	6a	5	5 15	1 91	Both Prem's
13c	52	1 18	1 10	11c	57½	1 14	1 09	6c	46	1 42	97	6c	59½	1 00	97	Negative

RANK DIFFICULTY AND DIAGNOSTIC VALUES OF TESTS—FIRST FORM

Part A				Part B				Part C				Part D				Remarks
Name	Rank culty	Diag Val Test	Diag Val Intel	Name	Rank culty	Diag Val Test	Diag Val Intel	Name	Rank culty	Diag Val Test	Diag Val Intel	Name	Rank culty	Diag Val Test	Diag Val Intel	
17b	15	3 50	1 66	15b	34	1 54	1 24	18b	20	2 63	1 45	12b	16	2 76	1 36	Both Prem's Negative
20c	58½	1 00	1 10	20c	60	1 27	1 23	5c	56	1 23	1 16	19c	55	1 03	1 14	
1c	22	2 36	1 23	7c	15	2 16	1 58	Negative Conclusions from Positive Premises				17c	17	2 95	1 42	Undistrib- uted Mid- dle
2c	5	2 70	1 40	2c	39½	1 13	.86	13c	37	1 45	.97	8c	50½	1 10	1 23	
6b	38½	1 46	1 16	6b	47½	1 32	1 03	9b	46	1 16	1 03	7b	46	1 22	1 23	Illhet (Major)
12c	27	2 00	.94	12c	41	1 16	1 16	7c	40½	1 05	.97	—	—	—	—	
14b	19½	2 50	1 44	8b	37	1 69	1 06	15b	28	2 29	1 18	5b	38	1 67	1 23	Same
18c	4	2 00	1 24	17c	6	7 00	2 38	17c	13	3 50	1 08	18c	20½	2 13	1 48	
19c	2	5 00	1 60	19c	18	2 74	2 00	16c	16	2 51	1 28	15c	15	2 62	1 46	False Over- sion
1a	54½	1 10	1 03	7a	53	1 13	.94	Universal Conclusions from a Particular Premise				17a	48	1 21	1 25	
18b	58½	1 00	.97	17b	52	1 03	1 22	20a	50	1 16	1 30	18b	40	1 70	1 48	Undis dle
19a	47	1 25	1 17	19a	49	1 40	1 50	16a	60	.97	.97	15a	56	1 00	1 06	
17a	21	2 80	1 40	Conclusion from Two Particular Premises				17a	48	1 21	1 25	Undis				Undis dle
17b	15	3 50	1 66	15a	23	2 46	1 82	1a	18	3 20	1 09	12a	27½	1 93	1 28	
17c	17	3 18	1 37	15b	34	1 54	1 24	1b	33	2 04	1 58	12b	16	2 76	1 36	
				15c	16	3 10	1 87	1c	24	2 15	1 03	12c	36	1 93	1 13	

RANK DIFFICULTY AND DIAGNOSTIC VALUES FOR EACH ITEM OF TEST—SECOND FORM

Part A					Part B					Part C					Part D					Remarks
Name	Rank	Diag. Val	Diag. for		Name	Rank	Diag. Val	Diag. for		Name	Rank	Diag. Val	Diag. for		Name	Rank	Diag. Val	Diag. for		
			cultry	Intel				cultry	Intel				cultry	Intel				cultry	Intel	
Valid Conclusions																				
2a	13½	1 36	1 64		9a	42½	1 79	94		8a	45½	1 50	1 09		7a	47	1 60	1 13		
2b	35½	1 26	1 13		9b	34½	2 12	1 50		8b	34½	2 00	1 41		7b	28	2 25	1 67		
2c	4	1 85	1 44		9c	42½	1 79	1 10		8c	45½	1 50	1 24		7c	44	1 51	1 28	Some not	
4b	31½	1 16	1 27		2b	31½	7 9	1 20		5b	27½	1 69	88		8b	45	1 08	1 20	Some not	
6c	54	1 05	1 00		7c	36	1 39	1 36		3c	39½	1 75	1 20		2c	31	1 28	1 20		
7a	50	90	1 06		5a	52½	93	1 06		2a	53½	97	1 13		4a	58	1 10	1 00		
8a	39	1 29	1 13		4a	29½	1 25	1 13		6a	31½	1 23	1 06		3a	27	1 28	1 06		
8b	16	1 25	1 44		4b	40	1 63	1 06		6b	31½	1 33	88		3b	8	1 76	1 54		
9b	31½	1 43	1 27		11b	52½	93	1 13		16b	33	82	1 37		13b	54	3 10	1 41		
10a	48	90	1 00		10a	26½	1 02	1 44		12a	36½	1 00	1 37		11a	52	1 32	1 32	Some not	
12a	6	1 69	1 20		13a	25	1 37	1 46		18a	24	1 52	1 39		14a	5	2 20	1 33	Some not	
12b	54	1 05	1 00		13b	42½	1 05	1 00		18b	42	92	1 21		14b	43	1 41	1 16		
13c	8	1 00	78		15c	56½	1 24	88		9c	42	1 33	1 36		15c	42	1 33	1 10		
14b	9	1 75	1 03		14b	1	1 40	1 17		13b	16	1 38	2 15		12b	57	1 16	1 00	Some not	
18c	33	1 34	1 17		19c	14	87	82		20c	24	90	1 00		20c	34	1 74	1 38	Some not	
19b	46	1 18	1 10		17b	45	81	1 07		11b	45½	1 23	91		17b	51	1 18	1 14		
19c	30	1 64	1 33		17c	46	79	87		11c	30	1 33	1 24		17c	36½	1 24	1 14		
Undistributed Middle Term																				
16a	35½	1 41	1 41		18a	14	3 42	1 36		14a	12½	2 19	1 42		18a	16	3 22	1 76		
16b	25	2 16	1 56		18b	17	2 26	1 36		14b	16	1 79	1 15		18b	21	2 08	1 33		
16c	23	2 20	1 50		18c	17	1 95	1 16		14c	10½	2 04	1 32		18c	11	3 44	1 95		
3a	22½	2 00	1 36		8a	4	2 16	1 54		1a	9	5 40	1 64		1a	1	3 46	1 35	2 Parties.	
3b	22½	2 25	1 44		8b	6	2 71	1 60		1b	19	3 28	1 64		1b	9	2 80	1 45	2 Parties.	
3c	16	2 54	1 53		8c	26½	2 76	1 38		1c	26	3 76	1 53		1c	18	3 00	1 64	2 Parties.	
15a	35½	1 57	1 37		6a	29½	1 73	1 06		15a	27½	2 00	1 80		5a	6	3 20	1 62	2 Parties.	
15b	3	3 52	1 83		6b	22½	2 30	1 44		15b	24	2 14	1 24		5b	32	1 86	1 35	2 Parties.	
15c	22½	2 16	1 13		6c	3	2 04	1 36		15c	1½	2 32	96		5c	18	1 86	1 35	2 Parties.	

RANK DIFFICULTY AND DIAGNOSTIC VALUES FOR EACH ITEM OF TEST—SECOND FORM

Part A				Part B				Part C				Part D				Remarks
Name	Rank Diffy- cutty	Diag Val. Intel		Name	Rank Diffy- cutty	Diag Val. Intel		Name	Rank Diffy- cutty	Diag Val. Intel		Name	Rank Diffy- cutty	Diag Val. Intel		
		Test	Intel			Test	Intel			Test	Intel			Test	Intel	
Illust Process of Minor Term																
7b	58½	1 00	1 00	5b	47½	1 13	1 28	2b	55½	1 14	94	4b	59	1 05	1 00	Universal from P
10b	58½	1 00	1 00	10b	34½	2 30	1 44	12b	21	2 00	1 06	11b	48	1 26	1 16	Universal from P
11b	28½	1 89	1 64	12b	10½	2 95	1 20	10b	4	2 62	91	10b	4	1 82	1 06	
18a	7	2 75	1 33	19a	12	3 00	1 34	20a	1½	2 45	1 91	20a	25½	2 08	1 43	
Illust Process of Major Term																
4a	48	1 16	1 06	2a	8	2 42	1 36	5a	10½	3 28	1 64	8a	36½	1 32	1 13	
7c	28½	1 88	1 64	5c	17	3 38	1 54	2c	21	4 15	1 27	4c	22	2 76	1 50	Neg fr pos
10c	39	1 44	1 36	10c	8	1 11	91	12c	7	1 40	1 06	11c	24	1 69	1 56	Some not
11a	13½	2 40	1 64	12a	2	2 59	1 36	10a	6	1 71	97	10a	3	2 48	1 13	
11c	26	2 00	1 13	12c	8	1 55	83	10c	5	1 17	91	10c	14	1 55	89	Some not
14a	18	1 57	83	14a	14	2 22	1 95	13a	8	1 56	85	12a	29	2 04	1 03	Some not
19a	1	2 08	1 10	17a	19	2 72	1 80	11a	12½	3 47	1 60	17a	2	2 57	1 26	Neg fr. pos
Illust Conclusion																
6a	48	1 05	94	7a	31½	2 04	1 75	3a	42	1 85	1 06	2a	20	3 20	1 64	
13c	8	1 00	78	15c	28	1 24	88	9c	42	1 33	1 36	15c	42	1 33	1 10	
20a	12	3 06	2 47	20a	5	1 27	1 78	19a	3	1 78	1 50	19a	30	2 18	1 67	Some not
Conclusion Drawn from Two Negative Premises																
5a	9	3 60	1 64	3a	60	97	1 00	4a	57½	1 08	1 00	6a	39	2 16	1 40	Pos fr neg
5b	54	1 05	.94	3b	39	1 72	1 75	4b	39½	1 85	1 27	6b	49	1 42	.91	
5c	54	1 05	1 06	3c	58½	97	94	4c	59	1 03	1 06	6c	40	2 07	1 50	
17a	20	2 26	1 21	16a	20	4 00	1 90	17a	16	3 83	1 67	16a	33	2 26	1 48	
17b	41	1 43	1 54	16b	38	1 80	1 52	17b	38	1 54	1 30	16b	25½	2 42	1 70	
17c	27	2 04	1 18	16c	21	3 93	1 47	17c	21	3 60	1 55	16c	11	1 41	1 77	

RANK DIFFICULTY AND DIAGNOSTIC VALUES FOR EACH ITEM OF TEST—SECOND FORM

Part A				Part B				Part C				Part D				Remarks
Name	cultry	Rank	Diag Val for Test Intel	Name	cultry	Rank	Diag Val for Test Intel	Name	cultry	Rank	Diag Val for Test Intel	Name	cultry	Rank	Diag Val for Test Intel	
4c	2	2 72	1 00	2c	24	1 63	1 54	<i>Affirmative Conclusion from Negative Premises</i>								
5a	9	3 60	1 64	3a	60	97	1 00	5c	16	2 35	1 20	8c	17	1 82	1 75	
17b	41	1 43	1 54	16b	38	1 80	1 52	4a	57½	1 08	1 00	6a	39	2 16	1 40	2 neg prems
8c	58½	1 00	1 00	4c	52½	1 25	1 20	17b	38	1 54	1 30	16b	25½	2 42	1 70	
12c	54	95	1 00	13c	56½	93	1 10	6c	57½	1 08	88	3c	54	1 22	1 20	
14c	35½	1 57	1 41	14c	49	95	1 44	18c	51	96	1 33	14c	57	1 02	1 17	
15b	3	3 52	1 83	6b	22½	2 30	1 44	13c	49	1 30	1 40	12c	50	1 39	1 09	
18b	44	1 42	1 17	19b	55	1 00	1 22	15b	24	2 14	1 24	5b	32	1 86	1 35	
20b	58	1 15	1 19	20b	56½	82	1 08	20b	58½	81	1 00	20b	56	1 15	1 11	
								19b	51	78	1 05	19b	60	1 00	1 15	Univ from partic
7c	28½	1 88	1 64	5c	17	3 38	1 54	<i>Negative Conclusion from Affirmative Premises</i>								
6b	16	2 54	1 75	7b	42½	1 73	1 20	2c	21	4 15	1 27	4c	22	2 76	1 50	Ill. Maj
19a	1	2 08	1 10	17a	19	2 72	1 80	3b	36½	2 22	1 36	2b	46	1 42	1 45	
								11a	12½	3 47	1 60	17a	2	2 57	1 26	Ill. Maj
7b	58½	1 00	1 00	5b	47½	1 13	1 28	<i>Universal Conclusion from a Particular Premise</i>								
10b	58½	1 00	1 00	10b	34½	2 30	1 44	2b	55½	1 14	94	4b	59	1 05	1 00	Ill Min.
20b	58½	1 15	1 19	20b	56½	82	1 08	12b	21	2 00	1 06	11b	48	1 26	1 16	Ill Min
								19b	51	78	1 05	19b	60	1 00	1 15	Pos fr neg
3a	22½	2 00	1 36	8a	4	2 16	1 50	<i>Conclusion from Two Particular Premises</i>								
3b	22½	2 25	1 44	8b	6	2 71	1 60	1a	9	5 40	1 64	1a	1	3 46	1 35	Undis. Mid.
3c	16	2 54	1 53	8c	26½	2 76	1 38	1b	19	3 28	1 64	1b	9	2 80	1 45	Undis. Mid.
15a	35½	1 57	1 37	6a	29½	1 73	1 06	1c	26	3 76	1 53	1c	18	3 00	1 64	Undis. Mid.
15b	3	3 52	1 83	6b	22½	2 30	1 44	15a	27½	2 00	1 80	5a	6	3 20	1 62	Undis. Mid.
15c	22½	2 16	1 13	6c	3	2 04	1 36	15b	24	2 14	1 24	5b	32	1 86	1 35	Undis. Mid.
								15c	1½	2 32	96	5c	18	1 86	1 35	Undis. Mid.

B. AVERAGE VALUES OF RANK DIFFICULTY, DIAGNOSTIC VALUE FOR THE TEST, AND DIAGNOSTIC VALUE FOR INTELLIGENCE GIVEN FOR EACH PART OF THE TEST AND FOR THE TEST AS A WHOLE BOTH FORMS INCLUDED

<i>Name of Fallacy</i>	<i>Part A</i>		<i>Part B</i>		<i>Part C</i>		<i>Part D</i>		<i>Whole</i>																					
Concl. from 2 Partic.	19	4	2	72	1	45	18	2	2	31	1	48	20	2	2	92	1	39	18	1	2	53	1	39	18	9	2	62	1	43
Undistrib Middle Term	22	2	2	36	1	36	18	5	2	63	1	54	18	2	2	88	1	32	18	5	2	71	1	48	19	3	2	64	1	42
Illicit Major Term.	26	0	1	82	1	18	16	4	1	96	1	26	15	4	2	08	1	11	21	9	2	01	1	21	19	9	1	97	1	19
Neg from Pos Premis.	16	3	2	45	1	35	28	2	2	50	1	46	26	0	2	66	1	20	25	7	1	84	1	36	24	0	2	36	1	34
Illicit Minor Term	36	7	1	56	1	20	18	8	2	88	1	57	16	1	2	68	1	38	28	0	2	01	1	34	24	9	2	28	1	37
Illicit Conversion.	23	1	2	37	1	39	20	8	2	01	1	63	31	5	1	82	1	23	30	4	2	03	1	37	26	4	2	06	1	40
Valid Conclusions	31	0	1	32	1	20	32	8	1	42	1	15	33	0	1	29	1	24	36	0	1	52	1	22	33	2	1	39	1	21
2 Negative Prem- ises.	31	7	1	94	1	25	37	3	2	08	1	48	35	9	1	90	1	21	32	3	2	20	1	31	34	3	2	03	1	31
Positive from Neg Premis.	33	7	2	03	1	35	48	9	1	26	1	22	44	4	1	45	1	14	36	6	2	15	1	31	40	9	1	72	1	26
Universal from Partic.	56	0	1	08	1	06	48	6	1	30	1	23	48	0	1	19	1	08	52	0	1	20	1	18	51	1	1	19	1	14

*c. Table Giving Each Item of Test as it Appeared in the Four
Different Kinds of Material*

In the following table the separate items of the test are given. Each item is given as it appears in the four different parts of the test, Parts A, B, C and D. They are given in the order in which they appeared in Part A. After each conclusion is given the per cent of subjects marking that item incorrectly, the rank difficulty derived from this, and the diagnostic value of that item for the test. It was impossible to group these according to fallacies as the three conclusions from one set of premises might represent three different fallacies.

The items of the first form are given first and are marked thus: A 1., B 3. etc. The items of the second form are marked thus: A* 1., B* 3.

FIRST FORM

A 1

Some of the boats on the river are sail-boats; Robert's boats are on the river;

therefore

a. Robert's boats are sail boats.	1.3%	R 54	1 10
b. some of Robert's boats are sail-boats	2.5%	R 13	2 21
c. some of Robert's boats are not sail-boats.	18.9%	R 22	2 36

B 7.

Some x's are y's; all z's are x's;

therefore

a. all z's are y's.	5%	R 53	1 13
b. some z's are y's.	73%	R 1	1 72
c. some z's are not y's.	32.8%	R 5	2 16

C 20.

Some arominia tupitandia are plastronomycetes; all rodomes are arominia tupitandia;

therefore

a. all rodomes are plastronomycetes.	5.5%	R 50	1 16
b. some rodomes are plastronomycetes.	72%	R 1	2 53
c. some rodomes are not plastronomycetes.	40%	R 0	4 80

D 17.

Some sweet-scented flowers are red; all roses are sweet-scented;

therefore

a. all roses are red.	53.5%	R 48	1 21
b. some roses are red.	44%	R 4	4 00
c. some roses are not red.	25.3%	R 17	2 95

A 2.

All the people living on this farm are related to the Joneses; these old men live on this farm;

therefore

a. these old men are related to the Joneses.	3%	R 49	1 05
b. all the people related to the Joneses are these old men.	1.3%	R 54	1 11
c. some people related to the Joneses are not these old men.	32.4%	R 5	2 70

B 2

All x's are y's; all z's are x's;

therefore

a. all z's are y's.	3.8%	R 55	.97
b. all y's are z's.	38%	R 9	2 78
c. some y's are not z's.	12.7%	R 39	1 13

C 13

All lysimachion is epilobium; all adenocaulon is lysimachion;

therefore

a. all adenocaulon is epilobium.	2.8%	R 58	.87
b. all epilobium is adenocaulon.	30.6%	R 21	2 71
c. some epilobium is not adenocaulon.	15%	R 37	1 45

D 8.

All Anglosaxons are English; all British are Anglosaxons;
therefore

a. all British are English	6.2%	R 47	.95
b. all English are British.	33.8%	R 11	3.70
c. some British are not English.	5%	R 50	1.10

A 3.

The school is west of the post-office; the court-house is west of the school;
therefore

a. the court-house is west of the post-office.	4.06%	R 46	1.11
b. the court-house is east of the post-office.	1.3%	R 54	1.11
c. the post-office is east of the court-house.	9.4%	R 36	1.11

B 5

x is west of y; z is west of x;
therefore

a. z is west of y.	7.6%	R 30	1.25
b. z is east of y.	2.5%	R 59	1.13
c. y is east of z.	9%	R 46	1.22

C 19

Crogingalong is west of Boggabilla; Pitarpunga is west of Crogingalong;
therefore

a. Pitarpunga is west of Boggabilla.	5.2%	R 51	.90
b. Pitarpunga is east of Boggabilla.	1.7%	R 59	1.07
c. Boggabilla is east of Pitarpunga.	8.7%	R 44	.90

D 13.

Chicago is west of San Francisco; Reno is west of Chicago;
therefore

a. Reno is west of San Francisco.	15.2%	R 36	1.25
b. Reno is east of San Francisco.	5.1%	R 49	1.26
c. San Francisco is east of Reno.	15.4%	R 34	1.11

A 4

The books on reference for the history course may not be withdrawn
from the library; the leather-bound books are not on reference for the
library course;

therefore

a. the leather-bound books may be withdrawn	36.4%	R 3	3.35
b. the leather-bound books may not be with- drawn.	3%	R 49	1.17
c. the books which may be withdrawn are not leather-bound.	13%	R 30	1.17

B 4.

A is not B; C is not A;
therefore

a. C is B.	3.8%	R 55	1.19
b. C is not B.	26.6%	R 19	2.96
c. B is not C.	25.3%	R 24	3.16

C 3

Heterophyllus are not zostera; epihydrus are not heterophyllus;
therefore

a. epihydrus are zostera.	6.2%	R 48	1.32
b. epihydrus are not zostera.	25%	R 26	2.76
c. zostera are not epihydrus.	20%	R 33	1.92

D 2.

No child under ten years of age is admitted to the movies unaccompanied by an adult; John is over ten years of age;
therefore

a. John will be admitted to the movies unaccompanied by an adult.	57.5%	R 1	4.70
b. John will not be admitted to the movies unaccompanied by an adult	5%	R 50	1.10
c. No boy admitted to the movies unaccompanied by an adult is John.	11.2%	R 42	1.58

A 5

If George is to the right of Sallie, and Marcus is to the left of George;
then

a. Marcus is at Sallie's left.	9.4%	R 36	1.55
b. Marcus is at Sallie's right.	27%	R 12	2.08
c. Sallie is at Marcus's right.	12%	R 32	1.74

B 3.

If x is to the right of y, and z is to the left of x;
then

a. z is at y's left.	12.7%	R 39	1.40
b. z is at y's right.	24%	R 27	2.60
c. y is at z's right.	8.8%	R 47	1.32

C 12.

If Congaree is to the right of Nemophilla, and Siscumbaba is to the left of Congaree;
then

a. Siscumbaba is to the left of Nemophilla.	13.5%	R 38	1.43
b. Siscumbaba is to the right of Nemophilla.	21.6%	R 30	1.60
c. Nemophilla is to the right of Siscumbaba.	12.3%	R 39	1.35

D 10.

If New York is north of Maryland, and North Carolina is south of New York;
then

a. North Carolina is south of Maryland.	17.7%	R 31	2.16
b. North Carolina is north of Maryland.	13.9%	R 39	1.47
c. Maryland is north of North Carolina.	15.2%	R 36	1.93

A 6.

All students of literature make use of reference libraries; all students of literature have intellectual curiosity;
therefore

a. some people who have intellectual curiosity make use of reference libraries	14%	R 28	1.55
b. some people who make use of reference libraries do not have intellectual curiosity.	8%	R 38	1.46
c. all people who have intellectual curiosity make use of reference libraries.	21%	R 18	1.85

B 6.

All a's are b's; all a's are c's;
therefore

a. some c's are b's.	36.6%	R 11	2.29
b. some b's are not c's.	8.8%	R 47	1.32
c. all c's are b's.	36.6%	R 11	3.93

C 9.

All foraminafera are rhyzopoda; all foraminafera are protozoa;
therefore

a. some protozoa are rhyzopoda.	25%	R 26	1.55
b. some rhyzopoda are not protozoa.	8.6%	R 46	1.16
c. all protozoa are rhyzopoda.	35.8%	R .5	3.70

D 7.

All professors in great universities are intellectual; all professors in
great universities are highly educated;
therefore

a. some highly educated people are intel- lectual.	20%	R 29	2.34
b. some highly educated people are not intel- lectual.	22.4%	R 22	2.34
c. all highly educated people are intellectual	7.5%	R 46	1.22

A 7.

None of the men employed by this company will be laid off this month;
all the men living in this house are employed by this company;
therefore

a. none of the men living in this house will be laid off this month.	2.7%	R 49	1.05
b. all the men living in this house will be laid off this month.	1.3%	R 54	1.11
c. some of the men living in this house will not be laid off this month.	51%	R 1	2.36

B 1.

No a's are b's; all c's are a's;
therefore

a. no c's are b's.	10.2%	R 43	1.4
b. all c's are b's.	6.3%	R 51	1.19
c. some c's are not b's.	43%	R 7	2.15

C 14.

No bactoringia are cytoplasts; all gitanjori are bactoringia;
therefore

a. no gitanjori are cytoplasts.	4%	R 54	.94
b. all gitanjori are cytoplasts.	4%	R 54	1.06
c. some gitanjori are not cytoplasts.	48%	R 6	1.80

D 1.

No nervous people are good dancers; all high schools girls are nervous;
therefore

a. no high high school girls are good dancers.	2.5%	R 53½	1.00
b. all high school girls are good dancers.	1.25%	R 57½	1.05
c. some high schools girls are not good dancers.	37.5%	R 8	2.33

A 8.

All seniors take History of Education; the Joneses take History of Educa-
tion;
therefore

a. the Joneses are seniors.	25.6%	R 13½	3.36
b. some of the Joneses are seniors.	17.5%	R 25	2.22
c. some of the seniors are Joneses.	31%	R 7	3.36

B 10.

All a's are b's; all c's are b's;
therefore

a. all c's are a's.	39%	R 8	3.93
b. some c's are a's.	26.6%	R 19½	1.82
c. some a's are c's.	26.4%	R 21½	1.69

C 10.

All clytia flavidula are hydromedusae; all obelia geniculata are hydromedusae;
therefore

a. all clytia flavidula are obelia geniculata.	43.6%	R 7	6.70
b. some clytia flavidula are obelia geniculata	41%	R 9	2.20
c. some obelia geniculata are clytia flavidula	36%	R 11	2.20

D 4.

All freshmen wear green caps; all these fellows wear green caps;
therefore

a. all these fellows are freshmen	38.8%	R 7	1.34
b. some of these fellows are freshmen	25%	R 18	2.08
c. some freshmen are among these fellows.	23.7%	R 19	1.96

A 9

Persons that have been educated in music enjoy symphony concerts;
the Smiths have not been educated in music;
therefore

a. the Smith's do not enjoy symphony concerts.	13%	R 30½	1.65
b. some of the people who enjoy symphony concerts are not Smiths.	17%	R 25	1.00
c. some of the Smiths do not enjoy symphony concerts.	20%	R 19½	1.96

B 14.

All x's are y's; no z's are x's;
therefore

a. no z's are y's.	45%	R 5	3.05
b. some y's are not z's.	49%	R 3	1.96
c. some z's are not y's.	34%	R 14	1.11

C 2.

All amphineura are mollusca; no scaphopoda are amphineura;
therefore

a. no scaphopoda are mollusca.	61.7%	R 2	2.80
b. some mollusca are not scaphopoda.	50%	R 4	1.42
c. some scaphopoda are not mollusca.	38.3%	R 12	1.10

D 11.

All monkeys have tails; human beings are not monkeys;
therefore

a. human beings have no tails.	30%	R 14	3.00
b. some creatures having tails are not human beings.	16.3%	R 33	1.58
c. some human beings do not have tails.	20%	R 29½	1.50

A 10.

None of the boys from this school went to the dance Saturday night; all the boys from this school are good dancers;
therefore

a. no good dancers went to the dance Saturday night.	5%	R 44	1.81
b. some good dancers did not go to the dance Saturday night.	4%	R 45	1.28
c. some people who went to the dance Saturday night are not good dancers.	17%	R 23	2.30

B 9.

No a's are b's; all a's are c's;
therefore

a. no c's are b's.	45.6%	R 4	3.65
b. some c's are not b's.	38%	R 9½	1.93
c. some b's are not c's.	35.5%	R 13	.92

C 8.

No triphilla are dracontia; all triphilla are arisaema;

a. no arisaema are dracontia.	56%	R 3	3.65
b. some arisaema are not dracontia.	38.7%	R 11	2.60
*c. some arisaema are dracontia.	5%	R 52	1.25

*These items are not like the corresponding items in Parts A and B.

D 20.

None of the boys in this room are maniacs; all the boys in this room are high school students;
therefore

a. no high school students are maniacs.	20.3%	R 26	2.14
b. some high school students are not maniacs.	17.2%	R 32	1.80
c. some high school students are maniacs.	8.6%	R 45	1.41

A 11.

None of Mary's cats are black; no black cats are in this house;
therefore

a. all cats in this house are Mary's.	6.7%	R 41½	1.38
b. no cats in this house are Mary's.	6.7%	R 41½	1.24
c. none of the cats outside of this house are Mary's.	2.7%	R 49½	1.17

B 13.

No a's are b's; no b's are c's;
therefore

a. all c's are a's.	2.6%	R 57½	1.14
b. no c's are a's.	23.4%	R 28	2.20
c. nothing that is not c is a.	9.3%	R 45	1.58

C 4.

No borgentillea are lepsici; no lepsici are gynrecchia;
therefore

a. all gynrecchia are borgentillea.	8.6%	R 46	1.42
b. no gynrecchia are borgentillea.	26.3%	R 23	2.15
c. nothing that is not gynrecchia is borgentillea.	3.66%	R 57	1.02

D 9.

No people interested in modern drama have failed to read this book; no people who have failed to read this book are actors;
therefore

a. all actors are interested in modern drama.	50%	R 3	4.70
b. no actors are interested in modern drama.	1.25%	R 57½	1.05
c. no people who are not actors are interested in modern drama.	10%	R 44	1.35

A 12.

All the poems on this page are written by a child nine years old; all the poems on this page have won prizes in the recent contest;
therefore

a. some poems that have won prizes in the recent contest are written by a child nine years old.	9%	R 35	1.43
---	----	------	------

b. all the poems that have won prizes in the recent contest are written by a child nine years old.	17%	R 25	1.55
c. some poems that have won prizes in the recent contest are not written by a child nine years old	15%	R 27	2.00

B 12.

All x's are z's; all x's are y's;
therefore

a. some y's are z's.	29.5%	R 17	2.24
b. all y's are z's.	26.4%	R 21½	3.34
c. some y's are not z's.	11.5%	R 41	1.16

C 7.

All lavauxia are onagrads; all lavauxia are oenothera;
therefore

a. some oenothera are onagrads.	27.2%	R 22	1.50
b. all oenothera are onagrads.	42%	R 8	3.20
c. some oenothera are not onagrads.	11.1%	R 40½	1.05

D 16.

*All Mongolians have slant eyes; the Chinese have slant eyes;
therefore

a. some Chinese are Mongolians.	34.2%	R 10	1.45
b. some Mongolians are not Chinese.	21.3%	R 24	2.12
c. the Chinese are Mongolians	52.5%	R 2	4.84

*This item does not correspond to the others.

A 13.

No good physician advertises his cures; Dr. J. does not advertise his cures;
therefore

a. Dr. J. is a good physician.	27.4%	R 11	3.56
b. Dr. J. is not a good physician.	0	R 58½	1.00
c. Dr. J. is the only good physician	1.4%	R 52	1.18

B 11.

No x's are y's; no z's are y's;
therefore

a. all z's are x's.	3.8%	R 54	1.11
b. no z's are x's.	19.4%	R 30	2.08
c. all x's are z's.	2.6%	R 57½	1.14

C 6.

No ctenophora possess nematocysts; no scyphozoa possess nematocysts;
therefore

a. all scyphozoa are ctenophora.	9.9%	R 43	1.50
b. no scyphozoa are ctenophora.	11.1%	R 40½	1.58
c. all ctenophora are scyphozoa.	8.6%	R 46	1.42

D 6.

No eminent man is influenced by trifles; Lincoln was not influenced by trifles;
therefore

a. he was an eminent man.	43.7%	R 5	5.15
b. he was not an eminent man.	0	R 59½	1.00
c. he was the only eminent man.	0	R 59½	1.00

A 14.

All freshmen take History 1;
therefore

a. all students taking History 1 are freshmen.	8%	R 38½	1.46
b. some students taking History 1 are not freshmen.	20%	R 19½	2.50
c. some students taking History 1 are freshmen.	12%	R 32½	1.55

B 8.

All x's are y's;
therefore

a. all y's are x's.	25.3%	R 24½	2.76
b. some y's are not x's.	14.1%	R 37	1.69
c. some y's are x's.	19.2%	R 31	1.93

C 15.

All anthozoa are coelenterates;
therefore

a. all coelenterates are anthozoa.	20%	R 33	1.88
b. some coelenterates are not anthozoa.	24%	R 28	2.29
c. some coelenterates are anthozoa.	10%	R 42	1.38

D 5.

All negroes of pure stock have kinky hair and dark skin;
therefore

a. all people that have kinky hair and dark skin are negroes.	22.4%	R 22½	2.34
b. some people that have kinky hair and dark skin are not negroes.	15%	R 38	1.67
c. some people that have kinky hair and dark skin are negroes.	11.2%	R 42	1.42

A 15.

None of the Smiths' money is invested in real estate; none of this money belongs to the Smiths;
therefore

a. none of this money is invested in real estate	5%	R 43	1.40
b. some of this money is not invested in real estate.	13%	R 29	2.00
c. no money invested in real estate is part of this money.	6%	R 40	1.32

B 18.

No x's are y's; no z's are x's;
therefore

a. no z's are y's.	19.1%	R 32	2.58
b. some z's are not y's.	13.4%	R 38	1.68
c. no y's are z's.	16.4%	R 36	2.20

C 18.

No caliphantoxia are gynsepia; no rhodomanthi are caliphantoxia;
therefore

a. no rhodomanthi are gynsepia.	21.6%	R 30½	1.61
b. some rhodomanthi are not gynsepia.	31%	R 20	2.63
c. no gynsepia are rhodomanthi.	17%	R 36	1.42

D 14.

No oranges are apples; no lemons are oranges;
therefore

a. no lemons are apples.	31.1%	R 12	2.66
b. some lemons are not apples.	20.8%	R 25	6.33
c. no apples are lemons.	31%	R 13	2.93

A 16.

All perfect specimens are in case A;
therefore

a. no specimen in case A is imperfect.	32%	R 6	3.80
b. no perfect specimen is in some case other than A.	23%	R 16	2.43
c. no imperfect specimens are in case A.	29%	R 9	3.24

B 16.

All A is B;
therefore

a. no B is something that is not A.	18.7%	R 33	2.26
b. no A is something that is not B.	10.7%	R 42	1.42
c. nothing that is not A is B.	21.3%	R 29	2.58

C 11.

All laboramati are gypsochromati;
therefore

a. no gypsochromati is something that is not laboramati.	26%	R 25	2.45
b. no laboramati is something that is not gypsochromati.	22%	R 29	.81
c. nothing that is not laboramati is gypsochromati.	19.5%	R 35	1.62

D 3.

All gentlemen have money;
therefore

a. no one without money is a person who is not a gentleman.	20.2%	R 27 $\frac{1}{2}$	1.55
b. no gentleman is without money	2.5%	R 53 $\frac{1}{2}$	1.10
c. no one who is not a gentleman has money.	11.2%	R 42	1.42

A 17.

Some of the chickens on the lawn are not Mr. Brown's; some of the chickens on the lawn are not properly fed;
therefore

a. some properly fed chickens are not Mr. Brown's.	21%	R 21	2.80
b. some of Mr. Brown's chickens are properly fed.	25%	R 15	3.50
c. some of Mr. Brown's chickens are not properly fed.	22%	R 17	3.18

B 15.

some a's are not b's; some a's are not c's;
therefore

a. some c's are not b's.	26.3%	R 23	2.46
b. some b's are c's.	18.4%	R 34	1.54
c. some b's are not c's.	29.6%	R 16	3.10

C 1.

Some bandigastices are not hydaxichantia; some bandigastices are not reptosantici;
therefore

a. some reptosantici are not hydaxichantia.	33.4%	R 18	3.20
b. some hydaxichantia are reptosantici.	20%	R 33	2.04
c. some hydaxichantia are not reptosantici.	26.2%	R 24	2.15

D 12.

Some books are not worth reading; some books are not allowed in circulation;
therefore

a. some books allowed in circulation are not worth reading.	20.2%	R 27½	1.93
b. some books worth reading are allowed in circulation.	28%	R 16	2.76
c. some books worth reading are not allowed in circulation.	15.2%	R 36	1.93

A 18.

All good ballet dancers have many years of training; some of the dancers in this musical comedy have many years of training;
therefore

a. some of the dancers in this musical comedy are good ballet dancers.	27%	R 10	2.60
b. all good ballet dancers are in this musical comedy.	0%	R 58½	1.12
c. some of the dancers in this musical comedy are not good ballet dancers.	36%	R 4	2.00

B 17.

All a's are b's; some c's are b's;
therefore

a. some c's are a's.	52%	R 2	6.36
b. all a's are c's.	5.4%	R 52	1.03
c. some c's are not a's.	44.5%	R 6	7.00

C 17.

All timbostera are lyricambia; some optobocacia are lyricambia;
therefore

a. some optobocacia are timbostera.	50%	R 5	3.90
b. all timbostera are optobocacia.	4.7%	R 53	1.14
c. some optobocacia are not timbostera.	38%	R 13	3.50

D 18.

All devout churchmen are regular in attendance at church; some very religious people are regular in attendance at church;
therefore

a. some very religious people are devout churchmen.	43.3%	R 6	5.16
b. all devout churchmen are very religious.	13.7%	R 40	1.70
c. some very religious people are not devout churchmen.	23.6%	R 30½	2.13

A 19.

Some of John's rare books are on this shelf;
therefore

a. all the books on this shelf are John's rare books.	2%	R 47	1.25
b. some of the books on this shelf are John's rare books.	0%	R 58½	1.12
c. some of the books on this shelf are not John's rare books.	39%	R 2	5.00

B 19.

Some x's are y's;
therefore

a. all y's are x's.	8.5%	R 49	1.40
b. some y's are x's.	10.1%	R 44	1.69
c. some y's are not x's.	28.2%	R 18	2.74

C 16.

Some porifera are hydrozoa;
therefore

a. all hydrozoa are porifera.	1.5%	R 60	.97
b. some hydrozoa are porifera.	5.9%	R 49	1.16
c. some hydrozoa are not porifera.	35%	R 16	2.51

D 15.

Some of Mary's dresses are blue;
therefore

a. all blue dresses are Mary's.	1.3%	R 56	1.00
b. some blue dresses are Mary's.	3.94%	R 52	1.00
c. some blue dresses are not Mary's.	29%	R 15	2.62

A 20.

All good dancers dance frequently; the men belonging to this frat do not dance frequently;
therefore

a. the men belonging to this frat are not good dancers.	11%	R 34	1.19
b. no good dancers belong to this frat.	30%	R 8	1.60
c. all good dancers belong to this frat.	0%	R 58½	1.00

B 20.

All a's are b's; no c's are b's;
therefore

a. no c's are a's.	16.7%	R 35	1.13
b. no a's are c's.	24.2%	R 26	1.54
c. all a's are c's.	1.5%	R 60	1.27

C 5.

All hlothuroidea are echinozoa; no echinoidea are echinozoa;
therefore

a. no echinoidea are hlothuroidea.	32.4%	R 19	1.02
b. no hlothuroidea are echinoidea.	34%	R 17	1.36
c. all hlothuroidea are echinoidea.	3.8%	R 56	1.23

D 19.

All people having a sense of rhythm care for music; good dancers do not care for music;
therefore

a. good dancers have no sense of rhythm.	23.6%	R 20½	.90
b. nobody having a sense of rhythm is a good dancer.	36.2%	R 9	1.48
c. all people having a sense of rhythm are good dancers.	1.4%	R 55	1.03

SECOND FORM

A* 1.

If the garden is to the right of the house; and the stable to the left of the garden;
then

a. the stable is to the left of the house.	8.7%	R 39	1.49
b. the stable is to right of the house.	23.7%	R 11	2.21
c. the house is to right of the stable.	7.5%	R 42½	1.40

B* 9.

If a is to the right of b, and c is to the left of a;
then

a. c is at b's left.	14%	R 42½	1.79
b. c is at b's right.	23%	R 34½	2.12
c. b is at c's right.	14%	R 42½	1.79

C 8.*

If Symlantobaga is to the right of Oscarsayoga; and Pattabantista is to the left of Symlantobaga;
then

a. Pattabantista is to the left of Oscarsayoga.	14%	R 45½	1.50
b. Pattabantista is to the right of Oscarsayoga.	22%	R 34½	2.00
c. Oscarsayoga is to the right of Pattabantista.	14%	R 45½	1.50

D 7.*

If New York is to the right of Detroit; and Chicago is to the left of New York;
then

a. Chicago is to the left of Detroit.	10.2%	R 47	1.60
b. Chicago is to the right of Detroit.	23%	R 23	2.25
c. Detroit is at Chicago's right.	11.5%	R 44	1.51

A 2.*

All the people taking part in the play are trained actors; no trained actors would accept so low a salary;
therefore

a. no person on so low a salary is taking part in the play.	22.2%	R 13½	1.36
b. no person taking part in the play is on so low a salary.	10%	R 35½	1.26
c. some people taking part in the play are not on so low a salary.	42.5%	R 4	1.85

B 1.*

All x's are y's; no y's are z's;
therefore

a. no z's are x's.	35%	R 22½	.88
b. no x's are z's.	6%	R 52½	1.02
c. some x's are not z's.	46%	R 10½	1.32

C 7.*

All cosmanthi are phacelia; no phacelia are hydrolea;
therefore

a. no hydrolea are cosmanthi.	29%	R 29	.71
b. no cosmanthi are hydrolea.	12%	R 48	1.19
c. some cosmanthi are not hydrolea.	37%	R 16	1.48

D 9.*

All wealthy men are stupid; no stupid men are college men;
therefore

a. no college men are wealthy.	18.8%	R 38	1.05
b. no wealthy men are college men.	6.2%	R 53	1.16
c. some wealthy men are not college men.	37.5%	R 12	2.34

A 3.*

Some people in the audience are laughing; some children are in the audience;
therefore

a. some children are laughing.	18.5%	R 22½	2.00
b. some of the people laughing are children.	18.5%	R 22½	2.25
c. some children are not laughing.	21%	R 16	2.54

B 8.

Some a's are b's; some c's are a's;
therefore

a. some c's are b's.	51%	R 4	2 16
b. some b's are c's.	49%	R 6	2 71
c. some c's are not b's.	29%	R 26½	2 76

*C** 1.

Some lansagrienses are hispanioli; some rhizcostodes are lansagrienses;
therefore

a. some rhizcostodes are hispanioli.	42%	R 9	5.40
b. some hispanioli are rhizcostodes.	36%	R 19	3.28
c. some rhizcostodes are not hispanioli.	32%	R 26	3 76

D 1.

Some dogs have long hair; some pet animals are dogs;
therefore

a. some pet animals have long hair.	52.5%	R 1	3.46
b. some long-haired animals are pets.	38 7%	R 9	2.80
c. some pet animals do not have long hair	27.5%	R 18	3 00

*A** 4.

All the flowers in this garden were grown from seed; some of the flowers
in this garden are not for sale;
therefore

a. some flowers that are for sale are not grown from seed.	3.7%	R 18	1.16
b. some flowers grown from seed are not for sale	11%	R 31½	1 16
c. some flowers grown from seed are for sale.	58%	R 2	2.72

*B** 2.

All x's are y's; some x's are not z's;
therefore

a. some z's are not y's.	47%	R 8	2 42
b. some y's are not z's.	25%	R 31½	.79
c. some y's are z's.	33%	R 24	1 63

*C** 5

All daghistans are lurians; some daghistans are not zakians;
therefore

a. some zakians are not lurians.	40%	R 10½	3.28
b. some lurians are not zakians.	30%	R 27½	.69
c. some lurians are zakians.	37%	R 16	2 35

*D** 8.

All the flowers in this garden are roses; some of the flowers in this
garden are not white;
therefore

a. some white flowers are not roses.	19%	R 36½	1 32
b. some roses are not white.	11.4%	R 45	1.08
c. some roses are white.	29.2%	R 17	1.82

*A** 5.

No good athlete lets himself get into poor physical condition; John does
not let himself get into poor physical condition;
therefore

a. John is a good athlete.	27%	R 9	3.60
b. John is not a good athlete.	1.2%	R 54	1 05
c. John is the only good athlete.	1.2%	R 54	1.05

B 3.*

No a's are b's; no c's are b's;
therefore

a. all c's are a's.	0%	R 60	.97
b. no c's are a's.	16%	R 39	1.72
c. all a's are c's.	2%	R 58½	.97

C 4.*

No juritobians are cantabilians; no catixianti are cantabilians;
therefore

a. all catixianti are juritobians.	2%	R 57½	1.08
b. no catixianti are juritobians.	16%	R 39½	1.85
c. all juritobians are catixianti.	1%	R 59	1.03

D 6.*

No giants have normal ductless glands; no dwarfs have normal ductless glands;
therefore

a. all dwarfs are giants.	17.7%	R 39	2.16
b. no dwarfs are giants.	8.7%	R 49	1.42
c. all giants are dwarfs.	17.5%	R 40	2.07

A 6.*

All the men belonging to the Athletic Club belong to this club;
therefore

a. all the men belonging to this club belong to the Athletic Club.	3.7%	R 48	1.05
b. some of the men belonging to this club do not belong to the Athletic Club.	21%	R 16	2.54
c. some of the men belonging to this club belong to the Athletic Club.	1.2%	R 54	1.05

B 7.*

All a's are b's;
therefore

a. all b's are a's.	25%	R 31½	2.04
b. some b's are not a's.	14%	R 42½	1.73
c. some b's are a's.	21%	R 36	1.39

C 3.*

All Ichnogobs are Rasmania;
therefore

a. all Rasmania are Ichnogobs.	15%	R 42	1.85
b. all Rasmania are not Ichnogobs.	20%	R 36½	2.22
c. some Rasmania are Ichnogobs.	16%	R 39½	1.75

D 2.*

All good students make high grades on the intelligence tests;
therefore

a. all those making high grades on the intelligence tests are good students.	26.3%	R 20	3.20
b. some of the people making high grades on the intelligence tests are not good students.	11.2%	R 46	1.42
c. some of the people making high grades on the intelligence tests are good students.	21.3%	R 31	1.28

A 7.*

All the rugs sold in that shop are high-priced; some of my rugs were sold in that shop;
therefore

a. some of my rugs are high-priced.	2.5%	R 50	.90
b. all high-priced rugs are mine.	0%	R 58½	1.00
c. some of my rugs are not high-priced.	14.8%	R 28½	1.88

B 5.*

All x's are y's; some z's are x's;
therefore

a. some z's are y's.	6%	R 52½	.93
b. all y's are z's.	9%	R 47½	1.13
c. some z's are not y's.	40%	R 17	3.38

C 2.*

All tigerlini are conferræ; some zoolidi are tigerlini;
therefore

a. some zoolidi are conferræ.	5%	R 53½	.97
b. all conferræ are zoolidi.	4%	R 55½	1.14
c. some zoolidi are not conferræ.	35%	R 21	4.15

D 4.*

All members of this club are undertakers; some florists are members of
this club;
therefore

a. some florists are undertakers.	2.5%	R 58	1.10
b. all undertakers are florists.	1.2%	R 59	1.05
c. some florists are not undertakers.	25.3%	R 22	2.78

A 8.*

No children who fight are allowed in this park; the children of the Jones
family are allowed in this park;
therefore

a. no children of the Jones family fight.	8.7%	R 39	1.29
b. no children who fight are children of the Jones family.	21%	R 16	1.25
c. all the children of the Jones family fight.	0%	R 58½	1.00

B 4.*

No a's are b's; all c's are b's;
therefore

a. no c's are a's.	27%	R 29½	1.25
b. no a's are c's.	15%	R 40	1.63
c. all c's are a's.	6%	R 52½	1.25

C 6.*

No pelomacholistræ are lipicanthræ; all pascolantia are lipicanthræ;
therefore

a. no pascolantia are pelomacholistræ.	25%	R 31½	1.23
b. no pelomacholistræ are pascolantia.	25%	R 31½	1.33
c. all pascolantia are pelomacholistræ.	2%	R 57½	1.08

D 3.*

No people who like to fight have much excess energy; boys have much
excess energy;
therefore

a. boys do not like to fight.	23.8%	R 27	1.28
b. no people who like to fight are boys.	38.8%	R 8	1.76
c. all boys like to fight.	5%	R 54	1.22

A 9.*

All of John's books are on this shelf;
therefore

a. no book on this shelf belongs to someone other than John.	7.5%	R 42	1.36
b. none of John's books are not on this shelf.	11%	R 31½	1.43
c. no book that is not John's is on this shelf.	6.2%	R 45	1.30

B 11.*

All x is y;
therefore

a. no y is something that is not x.	24%	R 33	1.92
b. no x is something that is not y.	6%	R 52½	.93
c. nothing that is not x is y.	20%	R 37	2.30

C 16.*

All vascocaloria are diorocura sapiens;
therefore

a. no diorocura sapiens is something other than vascocaloria.	22%	R 34½	1.52
b. no vascocaloria is something other than diorocura sapiens.	24%	R 33	.82
c. nothing other than vascocaloria is diorocura sapiens.	14%	R 45½	1.39

D 13.*

All well-trained musicians have good technique;
therefore

a. no musicians who have good technique are badly trained	38.4%	R 10	4.20
b. no musician with good technique is not well-trained.	36%	R 13	3.10
c. no musician who is not well-trained has good technique.	34.6%	R 15	2.54

A 10.*

None of Miss Smith's pupils are going to take part in the play; some of Miss Smith's pupils are very pretty;
therefore

a. some very pretty people are not going to take part in the play.	3.7%	R 48	.90
b. no very pretty people are going to take part in the play.	0%	R 58½	1.00
c. some people taking part in the play are not very pretty.	8.7%	R 39	1.44

B 10.*

No a's are b's; some a's are c's;
therefore

a. some c's are not b's.	29%	R 26½	1.02
b. no c's are b's.	23%	R 34½	2.30
c. some b's are not c's.	47%	R 8	1.11

C 12.*

No menimopyloria are gymnosargia; some menimopyloria are apodaceae;
therefore

a. some apodaceae are not gymnosargia.	20%	R 36½	1.00
b. no apodaceae are gymnosargia.	35%	R 21	2.00
c. some gymnosargia are not apodaceae.	48%	R 7	1.40

D 11.*

None of these books contain information concerning economics; some of these books are texts in economics;
therefore

a. some texts in economics do not contain information concerning economics.	6.3%	R 52	1.32
b. no texts in economics contain information concerning economics.	9.9%	R 48	1.26
c. some books containing information concerning economics are not texts in economics.	24.4%	R 24	1.69

A 11.*

Only artists are invited to the Kit-Kat Ball; my classmates are not invited to the Kit-Kat Ball;
therefore

a. my classmates are not artists.	22.5%	R 13½	2.40
b. no artists are classmates of mine.	14.8%	R 28½	1.89
c. some of my classmates are not artists.	16%	R 26	2.00

B 12.*

Only x's are y's; no z's are y's;
therefore

a. no z's are x's.	61%	R 2	2.59
b. no x's are z's.	46%	R 10½	2.95
c. some z's are not x's.	47%	R 8	1.55

C 10.*

Only calymbi myxomycetes are histoboli; rhaspodia chordata are not histoboli;
therefore

a. no rhaspodia chordata are calymbi myxomycetes.	52%	R 6	1.71
b. calymbi myxomycetes are not rhaspodia chordata.	56%	R 4	2.62
c. some rhaspodia chordata are not calymbi myxomycetes.	54%	R 5	1.17

D 10.*

Only masons and bricklayers belong to this union; plumbers do not belong to this union;
therefore

a. plumbers are not masons or bricklayers.	47.7%	R 3	2.48
b. masons and bricklayers are not plumbers.	41.7%	R 4	1.82
c. some plumbers are not masons or bricklayers.	35.4%	R 14	1.55

A 12.*

None of the boys that live in this block go to the public schools; all the members of my scout troop go to the public schools;
therefore

a. some of the members of my scout troop do not live in this block.	33%	R 6	1.69
b. none of the members of my scout troop live in this block.	1.2%	R 54	1.05
c. some of the boys living in this block are members of my scout troop.	1.2%	R 54	.95

B 13.*

No a's are b's; all c's are b's;
therefore

a. some c's are not a's.	32%	R 25	1.37
b. no c's are a's.	14%	R 42½	1.05
c. some a's are c's.	3%	R 56½	.93

C 13.*

No bolianthes are epineura; all banlox tabonaceae are epineura;
therefore

a. some banlox tabonaceae are not bolianthes.	34%	R 24	1.52
b. no banlox tabonaceae are bolianthes.	15%	R 42	.92
c. some bolianthes are banlox tabonaceae.	6%	R 51	.96

D 14.*

No corn is blue; all corn-flowers are blue;
therefore

a. some corn-flowers are not corn.	40%	R 5	2.20
b. no corn-flowers are corn.	14.3%	R 43	1.41
c. some corn is corn-flowers.	3.9%	R 57	1.02

A 13.*

No good children will be neglected by Santa Klaus;
therefore

a. all children who are not good will be neglected by Santa Klaus.	19.7%	R 19	2.40
b. all children who are not neglected by Santa Klaus are good.	37%	R 5	2.90
c. no children neglected by Santa Klaus are good.	31%	R 8	1.00

B 15.*

No x is y;
therefore

a. all that is not x is y.	9%	R 47½	1.11
b. all that is not y is x.	7%	R 50	1.00
c. no y is x.	28%	R 28	1.24

C 9.*

No antonikal sutoricci are vantinians;
therefore

a. all that are not antonikal sutoricci are vantinians.	4%	R 55½	1.14
b. all that are not vantinians are antonikal sutoricci.	6%	R 51	1.26
c. no vantinians are antonikal sutoricci.	15%	R 42	1.33

D 15.*

No respectable person would steal a dollar out of his neighbor's pocket;
therefore

a. anyone who is not respectable would steal a dollar out of his neighbor's pocket.	19.5%	R 35	2.20
b. anyone who would not steal a dollar out of his neighbor's pocket is respectable.	24.7%	R 23	2.20
c. no one who would steal a dollar out of his neighbor's pocket is respectable.	15.6%	R 42	1.33

A+ 14.

Some of those drawings are done by Dulac; none of Tom's collection are among those drawings;
therefore

a. some of Tom's collection are not done by Dulac.	20%	R 18	1.57
b. some drawings by Dulac are not in Tom's collection.	27.5%	R 9	1.75
c. some of Tom's collection are done by Dulac.	10%	R 35½	1.57

B 14.*

Some x's are y's; no z's are x's;
therefore

a. some z's are not y's.	42%	R 14	2.22
b. some y's are not z's.	63%	R 1	1.40
c. some z's are y's.	8%	R 49	.95

C 13.*

Some redautemplets are actinipellucidi; no gyrofantastices are redautemplets;
therefore

a. some gyrofantastices are not actinipellucidi.	47%	R 8	1.56
b. some actinipellucidi are not gyrofantastices.	37%	R 16	1.38
c. some gyrofantastices are actinipellucidi.	8%	R 49	1.30

D 12.*

Some drawings in red chalk are done by Michael Angelo; no good drawings are done in red chalk.
therefore

a. some good drawings are not done by Michael Angelo.	21.5%	R 29	2.04
b. some of Michael Angelo's drawings are not good.	3.7%	R 57	1.16
c. some good drawings are done by Michael Angelo.	7.6%	R 50	1.39

A 15.*

Most of the animals that live in the park are fed by the keepers; but many animals fed by keepers are not in good condition;
therefore

a. most of the animals that live in the park are not in good condition.	10%	R 35½	1.57
b. some animals in good condition live in the park.	51.3%	R 3	3.52
c. some animals in good condition do not live in the park.	19%	R 22½	2.16

B 6.*

Most x's are y's; but many y's are not z's;
therefore

a. most x's are not z's.	27%	R 29½	1.73
b. some z's are x's.	35%	R 22½	2.30
c. some z's are not x's.	56%	R 3	2.04

C 15.*

Most crinifolia are samposantes; but many samposantes are not corolia;
therefore

a. most crinifolia are not corolia.	30%	R 27½	2.00
b. some corolia are crinifolia.	34%	R 24	2.14
c. some corolia are not crinifolia.	63%	R 1½	2.32

D 5.*

Most fur coats are handsome; but many handsome coats are not cheap;
therefore

a. most fur coats are not cheap.	39.2%	R 6	3.20
b. some cheap coats are fur coats.	20.3%	R 32	1.86
c. some cheap coats are not fur coats.	27.5%	R 18	1.86

A 16.*

All the pink cups are Mary's; all the chipped cups are Mary's;
therefore

a. all the chipped cups are pink.	10%	R 35½	1.41
b. some of the chipped cups are pink.	16.5%	R 25	2.16
c. some of the pink cups are chipped.	17.5%	R 23	2.20

B 18.*

All x's are y's; all z's are y's;
therefore

a. all z's are x's.	42%	R 14	3.42
b. some z's are x's.	40%	R 17	2.26
c. some x's are z's.	40%	R 17	1.95

C 14.*

All evochtorides are nymphotoroides; all santoboli are nymphotoroides;
therefore

a. all santoboli are evochtorides.	38%	R 12½	2.19
b. some santoboli are evochtorides.	37%	R 16	1.79
c. some evochtorides are santoboli.	40%	R 10½	2.04

D 18.*

All followers of Tammany Hall are Democrats; all New York Irish-Americans are Democrats;
therefore

a. All New York Irish-Americans are followers of Tammany Hall.	32%	R 16	3.22
b. some of the New York Irish-Americans are followers of Tammany Hall.	25.4%	R 21	2.08
c. some followers of Tammany Hall are New York Irish-Americans.	38%	R 11	3.44

A 17.*

Some of his admirers are not really democratic; some of the people who are envious of him are not really democratic;
therefore

a. some of the people who are envious of him are not his admirers.	19.5%	R 20	2.26
b. some of the people who are envious of him are his admirers.	7.9%	R 41	1.43
c. some of his admirers are not envious of him.	15.8%	R 27	2.04

B 16.*

Some a's are not b's; some c's are not b's;
therefore

a. some c's are not a's.	37%	R 20	4.00
b. some c's are a's.	19%	R 38	1.80
c. some a's are not c's.	36%	R 21	3.93

C 17.*

Some ritascocolia are not cistamylopses; some ramellichopsia are not cistamylopses;
therefore

a. some ramellichopsia are not ritascocolia.	37%	R 16	3.83
b. some ramellichopsia are ritascocolia.	18%	R 38	1.54
c. some ritascocolia are not ramellichopsia.	35%	R 21	3.00

D 16.*

Some educated people are not courteous; some foreigners are not courteous;
therefore

a. some foreigners are not educated people.	20%	R 33	2.26
b. some foreigners are educated people.	24%	R 25½	2.42
c. some educated people are not foreigners.	15.8%	R 41	1.44

A 18.*

None of the internes in this hospital have been to see the new play; everyone who has seen the new play thinks the author is a genius;
therefore

a. no one who thinks the author a genius is an interne in this hospital.	32%	R 7	2.75
b. Some of the people who think the author a genius are internes in this hospital.	6.4%	R 44	1.42
c. some of the people who think the author a genius are not internes in this hospital.	10.2%	R 33	1.34

B 19.*

No x's are y's; all y's are z's;
therefore

a. no z's are x's.	43%	R 12	3.00
b. some z's are x's.	5%	R 55	1.00
c. some z's are not x's.	42%	R 14	.87

C 20.*

No igorots are pelomyxa; all pelomyxa are catusa;
therefore

a. no catusa are igorots.	63%	R 1½	2.45
b. some catusa are igorots.	5%	R 53½	.81
c. some catusa are not igorots.	34%	R 24	.90

D 20.*

No brunettes live in this block; all the people living in this block are Swedes;
therefore

a. no Swedes are brunettes.	24%	R 25½	2.08
b. some Swedes are brunettes.	4.2%	R 56	1.15
c. some Swedes are not brunettes.	19.7%	R 34	1.74

A 19.*

Some of the men in this town are interested in the new club; all the men in this town are truly patriotic;
therefore

a. some truly patriotic people are not interested in the new club.	61%	R 1	2.08
b. some truly patriotic people are interested in the new club.	3.9%	R 46	1.18
c. some people interested in the new club are truly patriotic.	14.3%	R 30	1.64

B 17.*

Some a's are b's; all a's are c's;
therefore

a. some c's are not b's.	39%	R 19	2.72
b. some c's are b's.	13%	R 45	.81
c. some b's are c's.	10%	R 46	.79

C 11.*

Some boarddentates are otitic; all boarddentates are relictoferantes;
therefore

a. some relictoferantes are not otitic.	38%	R 12½	3.47
b. some relictoferantes are otitic.	14%	R 45½	1.23
c. some otities are relictoferantes.	27%	R 30	1.33

D 17.*

Some of the girls in the chorus wear their hair braided; all the girls in the chorus have their hair bobbed;
therefore

a. some girls that have their hair bobbed do not wear their hair braided.	52%	R 2	2.57
b. some girls that have their hair bobbed wear their hair braided.	6.7%	R 51	1.18
c. some girls that wear their hair braided have their hair bobbed.	19%	R 36½	1.24

A 20.*

Some brown stockings are not darned;
therefore

a. some darned stockings are not brown.	23.6%	R 12	3.06
b. all darned stockings are brown.	0%	R 58½	1.15
c. none of the darned stockings are brown.	1.4%	R 51	1.26

B 20.*

Some x is not y;
therefore

a. some y is not x.	53%	R 5	1.27
b. all y is x.	3%	R .56½	.82
c. no y is x.	2%	R 58½	.82

C 19.*

Some pelmundoa are not cystidea;
therefore

a. some cystidea are not pelmundoa.	58%	R 3	1.78
b. all cystidea are pelmundoa.	6%	R 51	.78
c. no cystidea are pelmundoa.	0%	R 60	.69

D 19.*

Some of the Smiths are not at this party;
therefore

a. some of the people at this party are not Smiths.	21.4%	R 30	2.18
b. all the people at this party are Smiths.	0%	R 60	1.00
c. none of the people at this party are Smiths.	4.3%	R 55	.94

3. CONCLUSIONS

1. The fallacies most difficult for these subjects to detect were conclusions drawn from two particular premises and the undistributed middle term. The average rank difficulty of the items involving the first fallacy is 18.9. The average rank difficulty of the items involving the second is 19.3. The first fallacy violates the rule of the syllogism that states,³ "From two particular premises no conclusion can be drawn." Whenever this rule is violated there results an undistributed middle term, an illicit major or minor term, or else a conclusion from two negative premises. In the cases of this fallacy occurring in this test, the undistributed middle term results. It is interesting to observe that the fallacy of undistributed middle term is no easier to detect when it is involved with the fallacy of two particular premises. One might expect that both premises beginning with "some" would put the subjects on their guard, but apparently this is not the case. In one of these cases, 15 a, b, and c on the second form, the words "most" and "many" were used instead of "some," but this does not seem to have increased the difficulty.

2. The fallacy of undistributed middle term is difficult even when expressed in familiar material, though not quite as difficult as when expressed in symbolic or unfamiliar terms. There are two cases of this fallacy that proved easy to detect, 1 a and 18 b on the first form, but in these cases the fallacy of universal conclusion from a particular premise was involved and, since this latter fallacy proved in all cases to be easy to detect, it is undoubtedly *that* fallacy which was perceived by the subjects. In this fallacy of the undistributed middle term,³ "the middle term, the one used as a standard of comparison, is not used in either premise in its universal extent; thus we might be comparing the major term with one part of it and the minor term with another part. Such a comparison would, of course, not warrant us in either affirming or denying the connection of these terms in the conclusion." Item 8 a is a simple illustration.

All seniors take history of education; the Jones take history of education; therefore, the Jones are seniors.

3. The fallacy of Illicit Process of the Major Term ranks next in difficulty, the average rank difficulty of all these items

being 19.9. This fallacy is decidedly more difficult to perceive when expressed in symbolic or unfamiliar material than when expressed in familiar material. Its average rank in Part B is 16.6, in Part C 15.4, while in Part A its average rank is 26 and in Part B 21.9. Two cases of this fallacy, items 6 b and 12 c in the first form, were somewhat easier than the others, especially in Parts B and C. Another fallacy was involved in these two cases and this may have reduced the difficulty. This fallacy violates that rule of the syllogism that says,⁸ "No term must be distributed in the conclusion which was not distributed in one of the premises. That is, the conclusion must be proved by means of the premises and no term which was not employed in its universal signification in the premises can, therefore, be used universally in the conclusion." In other words, if the premises tell us only about a part of a class represented by the term, we cannot draw conclusions about the whole of that class. An illustration of this fallacy is item 10 c in the second form. This is somewhat interesting because it proved to be so very much more difficult in Parts B and C than in Part A. Only 8.7 per cent of the subjects fail to perceive this fallacy in Part A, but 47 per cent fail to perceive it in Part B and 48 per cent in Part C. The item is as follows:

None of Miss Smith's pupils are going to take part in the play; some of Miss Smith's pupils are very pretty; therefore

a.

b.

c. Some people taking part in the play are not very pretty.

4. The fallacy of the Illicit Process of the Minor Term proves not to be as difficult as the Illicit Major. The average rank difficulty for this fallacy is 24.9, and, like the Illicit Major, it is much more difficult in Parts B and C than in Parts A and D. One case of this fallacy, item 10 a in the first form, has a rank difficulty of 4 in Part B and of 44 in Part A. When it appeared in familiar material only 5 per cent of subjects marked it incorrectly, but when it appeared in symbolic material 45.6 per cent marked it incorrectly. This item is as follows:

None of the boys from this school went to the dance Saturday night; all of the boys from this school are good dancers; therefore, no good dancers went to the dance Saturday night.

Two cases of this fallacy, items 7 b and 10 b of the second

form, were very easy, probably because another fallacy was also involved, a Universal Conclusion drawn from a Particular Premise. The fallacy of Illicit Minor is the same as Illicit Major except that in one the minor term, or subject of the conclusion, is involved and in the other the major term or predicate of the conclusion is involved. Item 6 c of the first form is an illustration.

All students of literature make use of reference libraries; all students of literature have intellectual curiosity; therefore, all people who have intellectual curiosity make use of reference libraries.

Twenty-one per cent of subjects mark this incorrectly in Part A, 36 per cent in Part B, 36 per cent in Part C and 22 per cent in Part D.

5. The fallacy ranking in difficulty between the fallacies of Illicit Major and Minor is that of drawing a negative conclusion from two affirmative premises. This violates the rule of the syllogism which says, "If both premises be affirmative, the conclusion must be affirmative."

Nearly all the cases of this fallacy occurring in this test involve other fallacies. This fallacy is often more difficult in Part A than in the other parts, the average rank difficulty in Part A being 16.3, while in the other parts the average ranges from 25.7 to 28.

6. The fallacy of illicit conversion ranks next in difficulty, its average rank difficulty being 26.4. Most of the cases of this fallacy are combined with obversion. The cases where only conversion is involved are rather illuminating. Twenty-six per cent of the subjects marked as valid "All those making high grades on the intelligence tests are good students," deduced from "All good students make high grades on the intelligence tests." Twenty-five per cent of the subjects marked as valid "All y's are x's," deduced from "All x's are y's." Twenty per cent marked as valid "All coelenterates are anthozoa," deduced from "All anthozoa are coelenterates." The same difficulty is at the basis of this fallacy and all those mentioned so far. The subjects do not realize that certain terms used in the premises are used only in a partial sense, not in their universal extent. Unless the word "some" is placed before the term, they consider that "all" is implied. Of course when the case is as self-evident as "All horses are animals,"

they realize that this does not mean "All animals are horses"; but when the actual facts of the case are unknown to them, many subjects take it for granted that the predicate is used in its universal extent. When the proposition is not only falsely converted but is also obverted, the perception of the fallacy is very difficult, whatever the material.

7. It is evident that the subjects for the most part were ignorant of the fact that no formal conclusion can be drawn from two negative premises. Many of the cases of this fallacy were easy and many were difficult. The three cases where a positive conclusion is drawn from two negative premises are very much more difficult when expressed in familiar material than in symbolical unfamiliar material. These cases have been discussed above.

8. Except in these three cases, the subject found it easy to perceive that a positive conclusion could not be drawn when one or more of the premises was negative. The average rank difficulty of all such items is 40.9.

9. Where a universal conclusion was drawn from a particular premise, very few subjects fail to see the fallacy. This was the easiest fallacy for them to perceive. The average rank difficulty for all items was 51.1.

10. Many subjects seem to believe that the statement "Some are" means necessarily that "Some are not" and that "Some are not" means necessarily that "Some are." Therefore, if the premises are universal, these subjects think it invalid to draw a particular conclusion. That is, if the premise is "All are" these subjects think it invalid to conclude that "Some are" and if the premise states "None are" they think it invalid to deduce "Some are not" For instance item 7 c in Part A was marked valid by 51 per cent of the subjects.

7. None of the men employed by this company will be laid off this month; all the men living in this house are employed by this company; therefore,

- a.
- b.
- c. some of the men living in this house will not be laid off this month.

In every day life a statement about "some" does often imply that "some are not" especially when the "some" is accented.

11. "Some" in conclusions makes for difficulty. The average rank difficulty of all such conclusions is 24.17. The particular conclusions which are also negative are slightly more difficult than the affirmative ones, the average rank difficulty for the negative being 21.7 and the affirmative 27.9. The difficulty of the particular conclusions does not seem to be much affected by whether "some" appears in the premises or not, the average rank difficulty of the two groups being 22.87 and 25.8.

12. Negative conclusions where one or both premises are negative have an average rank difficulty of 30.1. Affirmative conclusions from negative premises are easier, their average rank difficulty being 33.7. Negative conclusions from affirmative premises are very difficult, the average rank difficulty being 16.3.

General Conclusion from This Part of the Study

Many college students are not on their guard against some of the simplest and commonest fallacies of thought. They do not have the aid of such simple general principles as that no conclusion can be drawn from two negative premises. They do not know that the converse of a true proposition that is universal and affirmative is not necessarily true. It would seem that for practical purposes it would be well worth while to give students some drill in such very simple and common fallacies so that they would at least be on their guard against them.

Diagnostic Value of Each Item and Each Kind of Fallacy for Ability in the Syllogistic Reasoning Test

It seemed worth while to see what part each item and each fallacy played in distinguishing the subjects who did well in the test from those subjects that did poorly. For anyone desiring to test students' ability along the lines of this sort of reasoning, it would be worth while to know which fallacies should be stressed and which might be slighted. For instance, in this test, very few cases of the fallacy of Illicit Process of the Minor Term were used, and yet these items proved to be extremely valuable in distinguishing the upper half from the lower half in ability on the test.

NOTE. Conclusions 11 and 12 are based on data from Part A alone.

A very crude and simple method was used for determining this diagnostic value. A four fold table was made for each item, having in the upper left hand corner the number of subjects scoring above the median in the test who marked that particular item correctly. In the upper right hand corner was placed the number of subjects scoring below the median on the test who marked that item correctly. In the lower left hand corner was placed the number of subjects scoring above the median in the test who marked that item incorrectly, and in the lower right hand corner was placed the number of those scoring below the median who marked that item incorrectly. Thus the larger the numbers in the upper left and lower right compared to the numbers in the upper right and lower left, the more valuable the item. Thus the best possible diagnostic value would be for all the subjects above the median to mark the item correctly, and all below the median to mark it incorrectly. No item in this test reached this maximum value. The item having the highest diagnostic value was 17 c. Its

four-fold table was like this— $\frac{34}{3} \mid \frac{6}{29}$, and reads,—Of the

subjects scoring above the median, 34 marked this item correctly and 3 marked it incorrectly. Of the subjects scoring below the median, 6 marked this item correctly and 29 marked it incorrectly. As for purposes of comparison, it was desirable to have this value expressed in terms of one number, this was obtained by dividing the sum of the numbers in the upper left and lower right hand corners by the sum of the numbers in the upper right and lower left hand corners. These figures are listed in the table in the column headed "Diag. value for test." In order that the reader might read these figures more understandingly, a table of the distribution of these values is given for each part of the test and the class-intervals are translated into terms of correlation coefficients obtained from the four-fold table by using Sheppard's Method of Unlike Signs.¹(10) Thus it is clear that a diagnostic value of 1.00 means no diagnostic value, that less than 1.00 means a negative diagnostic value; that the subjects good on the test were less successful in marking that item than the subjects poor on the test; that diagnostic values begin to have real sig-

¹ Rugg; Statistical Methods, Page 297.

nificance when they are 1.50 or above, and that diagnostic values of 3.00 or above are very good.

This does not pretend to be an accurate measure of the diagnostic value of the items, as it would not take into account the value of distinguishing the extremely good on the test from the moderately good, or the very poor from the moderately poor.

A glance at the table shows at once that many of the items have no diagnostic value because they are too easy. If all the subjects except one or two marked the item correctly, it might just as well be omitted from the test except for its value as an encouragement to the subject.

DISTRIBUTION OF DIAGNOSTIC VALUES FOR SYLLOGISM TEST.

<i>Diag. Val.</i>	<i>Coef. of Cor.</i>	<i>Frequencies</i>			
		<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>
Below 1.00		0	2	6	2
1.00-1.49	0-.31	28	21	22	24
1.50-1.99	.31-.51	9	12	10	11
2.00-2.49	.51-.63	10	9	7	9
2.50-2.99	.63-.70	4	7	7	5
3.00-3.49	.70-.76	5	4	2	1
3.50-3.99	.76-.80	3	3	4	1
4.00-4.49		0	0	0	2
4.50-4.99		0	0	1	3
5.00-5.49	.85-	1	0	0	2
5.50-5.99		0	0	0	0
6.00-6.49	.90-	0	1	0	0
6.50-6.99		0	0	1	0
7.00-7.49	.923-	0	1	0	0

DISTRIBUTION OF DIAGNOSTIC VALUES FOR INTELLIGENCE

<i>Diag. Val.</i>	<i>Coef. of Cor.</i>	<i>Frequencies</i>			
		<i>Part A</i>	<i>Part B</i>	<i>Part C</i>	<i>Part D</i>
Below 1.00		11	9	15	7
1.00-1.24	0-.16	27	18	26	21
1.25-1.49	.16-.31	12	9	10	20
1.50-1.74	.31-.41	8	10	5	6
1.75-1.99	.41-.50	0	7	3	4
2.00-2.24	.51-.60	1	5	0	1
2.25-2.49	.60-.63	1	2	1	0
2.50-2.99	.63-.70	0	0	0	1

Conclusions

Of the different fallacies, those of the Undistributed Middle Term and Conclusion drawn from Two Particular Premises have much the greatest diagnostic value for the test. Their average values for the whole test (both forms) are 2.64 and 2.62. Their diagnostic value remains high throughout the four parts of the test. The next most valuable in this connection are the fallacies of drawing a negative conclusion from affirmative premises, which has an average value of 2.36 and the Illicit Process of the Minor Term, which has an average value of 2.28. This latter fallacy is not so valuable in familiar material as in symbolic and unfamiliar. The fallacy of Illicit Major Term, though more difficult than these two fallacies, has not so high a diagnostic value.

The fallacy of Illicit Conversion is more valuable when obversion is also involved. The fallacy of a conclusion drawn from two negative premises is also fairly valuable, having an average value of 2.03.

The fallacy of Universal Conclusion from Particular Premise was so easy to detect that it had no diagnostic value. The valid conclusions, also, are of little diagnostic value, the most valuable items being those that are negative as well as particular.

To sum up, for tests of this kind the following fallacies should be stressed, as they have the greatest diagnostic value for the test; Undistributed Middle Term, Conclusion from Two Particular Premises, Illicit Minor Term and Negative Conclusions from Affirmative Premises.

Correlation of Each Item with Intelligence

An effort was made to discover which items and thence which fallacies were of most importance in correlating with intelligence. The subjects were divided into two groups according to their scores on the Thorndike Intelligence examination, one group with scores above the median, the other with scores below the median. The same method was used as was described above for finding the diagnostic value for the test. A four-fold table was made for each item, with the upper and lower left hand corners containing the numbers of those above the median in the intelligence test marking the item respectively correctly and incorrectly and with the upper and

lower right hand corners containing the number of those below the median in the intelligence test marking the item respectively correctly and incorrectly.

Conclusions

Naturally, since the test scores do not show a close relationship with intelligence, the items do not. The items of Parts B and D show decidedly more correlation than the items of Parts A and C. The items falling under one particular fallacy do not show as great consistency in correlation with intelligence as with the test itself. However, it is clear that the same fallacies that led in diagnostic value for the test also lead in diagnostic value for intelligence. The fallacies of Undistributed Middle Term, Conclusion from Two Particular Premises, Illicit Conversion and Illicit Minor Term head the list. The fallacy of Negative Conclusion from Affirmative Premises does not stand quite so high in diagnostic value for intelligence as for the test. The three cases of drawing an affirmative conclusion from two negative premises which were so much more difficult in Parts A and D than in Parts B and C show good correlation with intelligence in Parts A and D. In general, the fallacies of Universal Conclusion from Particular Premises, Affirmative Conclusion from Negative Premises, and Illicit Major, together with the valid conclusions, show little or no correlation with intelligence.

CHAPTER IV

Summary of Conclusions

1. Ability to do formal syllogistic reasoning is much affected by a change in the material reasoned about. The easiest material is the familiar and concrete. The most difficult is the unfamiliar (long words). The symbolic material is almost as difficult as the unfamiliar. The suggestive material is more difficult than the familiar but not so difficult as the symbolic and the unfamiliar.

2. The standings of individuals in tests of syllogistic reasoning are affected to some extent by changing the material reasoned about. The correlations between one part of the test and another are positive and high, but not so high as between different forms of the same part. The standings of certain individuals are very little affected by the change in material, whereas the standings of others are changed very materially.

3. There is a marked correlation, though not a high one, between success on the syllogism test and success in the Thorndike Intelligence Examination. The correlation between the Thorndike Intelligence Examination and success with the symbolic material is decidedly higher than with any of the other kinds of material used.

4. When the separate items are examined, it is seen that the most difficult fallacies are those of the Undistributed Middle Term, and Conclusion drawn from Two Particular Premises. Other difficult fallacies are Illicit Major and Minor Terms and Negative Conclusion from Affirmative Premises. Sometimes difficult but not consistently so is the fallacy of drawing a conclusion from two negative premises. Valid conclusions are often easy and universal conclusions from a particular premise are always easy.

5. Some items vary widely in their relative difficulty as the material is changed. Most items increase in difficulty as the material is changed from familiar to symbolic, etc., but a few items representing very common fallacies are much less difficult in symbolic material than in familiar. This is probably due to bad habits of every-day reasoning which are much in

force in the familiar situation, but are not so influential when the material is symbolic or unfamiliar.

6. The fallacies having the most value in distinguishing the successful from the unsuccessful in a test of this kind are the Undistributed Middle Term, Conclusion from Two Particular Premises, Negative Conclusion from Affirmative Premises and Illicit Minor. Many of the items were too easy to be of any diagnostic value for the test.

7. The same fallacies that have special value for determining success on the syllogism test have value for determining success on the Thorndike Intelligence Examination. The fallacy of Illicit Conversion has a slightly greater relative value for the diagnosis of intelligence than for the diagnosis of ability on the syllogism test.

BIBLIOGRAPHY

1. Arlht, Ada, and Hall, Margaret. Intelligence Tests versus Entrance Examinations as a Means of Predicting Success in College. *Journal of Applied Psychology*. 1923.
2. Bailor, E. M. Content and Form in Tests of Intelligence. *Columbia University Contributions to Education*. No. 162. 1924.
3. Creighton, J. E. *An Introductory Logic*. Macmillan Co. 1914.
4. Dunlap and Snyder. Practise Effects in Intelligence Tests. *Journal of Experimental Psychology*. 1920.
5. Garrett, H. E. *Statistics in Psychology and Education*. Longmans. 1926.
6. Jones, A. L. *Logic Inductive and Deductive*. Henry Holt and Co. 1909.
7. Kitson, H. D. *Scientific Study of College Students*. *Psychological Review Monograph*. No. 23.
8. McCall, W. A. *How to Measure in Education*. Macmillan Co. 1923.
9. Pintner, R. *Intelligence Testing*. Henry Holt and Co.
10. Rugg, H. O. *Statistical Methods Applied to Education*. Houghton Mifflin Co. 1917.
11. Thorndike, E. L. Effect of Changed Data on Reasoning. *Journal of Experimental Psychology*. 1922.
12. Thorndike, E. L. Practise Effects on Intelligence Tests. *Journal of Experimental Psychology*. 1922.
13. Wood, Ben D. *Measurement in Higher Education*. World Book Co. 1923.
14. Woodworth, R. S. *Psychology*. Henry Holt and Co. 1921.

VITA

Minna Cheves Wilkins, born August 6, 1884, in Ridgeland, South Carolina; A.B. Randolph-Macon Woman's College, 1905; A.M. Columbia University, 1916; Instructor in Psychology and Education, Randolph-Macon Woman's College, 1909-16; Statistician and Assistant in Mental Testing, Dept. of Psychology, Carnegie Institute of Technology, 1919-21; Associate Professor of Psychology, Hollins College, 1922-23; Psychologist, Juvenile Court, Cincinnati, Ohio, 1923-24; Psychologist, Dept. of Psychiatry, Vanderbilt Clinic, New York City, 1924-27; Psychologist, Mental Hygiene Clinic, State Charities Aid, Association for Improving the Condition of the Poor, and Brooklyn Bureau of Charities, New York City, 1927- .

REPRESENTATIVE CORRELATION TABLES

Score in Part A. Syllogism Test.

Score in Thorndike College Entrance Exam.	Score in Part A. Syllogism Test.										
	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100
105-119										2	
100-104									2	2	
95-99							2		3	3	
90-94			1					1	2	4	2
85-89	1			1			2	1	3	2	
80-84		1	2					1	1	2	2
75-79			2				2		3	2	
70-74				1	1	1	1	2			
65-69			1	1			1		2		
60-64		1		1			1				
55-59					1		1				
50-54											
45-49	1										

(F) Correlation Table showing scores in Part A of Syllogism Test and Scores in Thorndike College Entrance Exam.

Score on Part B - Syllogism Test.

Score on Thorndike College Entrance Exam.	Score on Part B - Syllogism Test.												
	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100
105-109												2	
100-104										1	1	2	
95-99									1	4	1	1	1
90-94	1						2	1	1	2		2	
85-89				1		1	2	1		1	3	1	
80-84			1		2				2	1	2	1	
75-79				2			1	2		2	1	1	
70-74				2		2	1		1				
65-69			1	2	1		1						
60-64		1				2							
55-59			1							1			
50-54													
45-49			1										

(9) Correlation Table showing scores in Part B of Syllogism Test and Scores in Thorndike College Entrance Exam.

Total Score First Form.

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Total Score - Second Form	95-99								1	7
90-94								3	3	3
85-89							1	3	7	1
80-84							6	5	1	
75-79						3	3	2		
70-74			1		1	2				
65-69		1	2	3	3	1				
60-64		2		2						
55-59	1	2		1						
50-54		5	1							

(H) Representative Correlation Table for Self Correlations.

Score on Part A.

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100
Score on Part B.										1	
100										1	
95-99							1		1	9	2
90-94							2	1	3	1	1
85-89					1		1		5	6	1
80-84								3	2	1	1
75-79							1		2	1	
70-74			1	1					3	1	
65-69	1	1		1	1		2	1	1		
60-64			3				1				
55-59			3	1		1	2		1		
50-54		2		1			1				
45-49				2							
40-44								1			

(I) Correlation Table showing Scores in Parts A and B of Syllogism Test.

THE EFFECT OF INCENTIVES ON ACCURACY OF DISCRIMINATION MEASURED ON THE GALTON BAR

BY
HUGHBERT C. HAMILTON, M. A.

Submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy, in the Faculty of Philosophy,
Columbia University.

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, Editor

No. 103

NEW YORK
March, 1929

The writer gladly acknowledges his indebtedness to Professor R. S. Woodworth for continued suggestion and aid from the very inception of the problem. Valued assistance was also given by Professors A. T. Poffenberger and H. E. Garrett.

TABLE OF CONTENTS

<i>Chapter</i>	<i>Page</i>
I. Historical Survey and Introduction	5
II. Experimental Situation	14
Apparatus	14
Procedure	20
III. Statement of Results	24
IV. Discussion of Results	28
V. On Adjusting One Length Equal to Another	40
VI. Summary and Conclusions	44
VII. Bibliography	46
VIII. Appendix	47

I. HISTORICAL SURVEY AND INTRODUCTION

An important focal point in a dynamic psychology is the study of motivation, particularly the nature of incentives and their effects. If prediction and control are the aim of the science of psychology, certainly here lies one of the keys to the goal. Many ostensibly different types of incentives have been applied to many different forms of behavior. Much of the observation has not taken the form of systematic experiment. And most of the studies which might deserve the latter term have been too poorly controlled or have employed too few subjects to permit of any reliable conclusions. However, there are summarized below, those studies in which relatively simple and specific external incentive stimuli have been applied to the more simple and accurately measurable performances, and studies of visual discrimination of length, it being the purpose of this investigation to study the effect of specific incentive stimuli upon the latter performance.

1. ANIMAL STUDIES

The not unequivocal distinction made between drive and incentive stimulus is useful here. As implied above, the determination of the conditions varying the strength of hunger, sex, and thirst drives, for instance, does not so much concern us as does the measuring of the effects of external stimuli acting as incentives. The latter has been attempted in receptor studies, for example.

Hoge and Stocking⁹ were interested in the relative value of punishment and reward. They used electric shock for punishment and food for reward. The behavior motivated was brightness discrimination, after the Yerkes method. Only two female rats, one albino and one black and white, constituted each of their three groups. One group received both punishment and reward, another was punished for wrong responses, and the third was rewarded for correct responses. Both rats receiving both punishment and reward learned, one rat receiving just punishment, and neither of the rats receiving just reward learned. This scarcely justifies the conclusion that the combined incentives are more effective than

punishment alone, and punishment is more effective than reward alone.

Dodson,³ in 1917, built up discrimination habits in white rats, also for the purpose of comparing reward and punishment. Shock was the stimulus for punishment, and food for reward. Punishment was given thus: if the rat went to the dark box, it received shock, and had to return to its nest through the light box; if it went to the light box, it simply passed on through to its nest. Reward was given thus: if the rat went to the light box it received food; if it went to the dark box, it had to return to its nest through the light box. Supposedly four degrees of punishment and of reward were used, i.e., four strengths of shock, and four degrees of hunger. For all degrees of intensity, punishment was more effective than reward. There was both an optimum strength of shock and an optimum degree of hunger, which were less than the highest intensities used.

Warden and Aylesworth¹⁸ tested the findings of Hoge and Stocking in respect to the comparison of reward and punishment by the discrimination method. They employed ten white rats in each of three groups. One group received only food for correct responses, another only shock for wrong responses, and the third received food for correct responses and shock for wrong responses. Combined reward and punishment proved to be most effective, punishment alone ranked next, and reward alone was least effective. In fact, the group receiving only food failed to learn, under the rather strict criteria employed.

2. STUDIES EMPLOYING HUMAN SUBJECTS

An early study which forms a sort of negative approach to the problem is that made by Judd.¹¹ He wanted to discover the effect of practice when the subject had no knowledge of the results of his performance. On a table was placed a line running at an angle from the subject. The table was divided in the center by a screen. The subject's task was to place a dot behind the screen at a point which would fall within the exact continuation of the line. Nine different angles were employed and each angle was tried twenty times on each of ten days. The one subject employed was unable to make improvement without any knowledge of how well or how poorly he was doing.

As late as 1923, Spencer¹⁵ repeated Judd's experiment, using four subjects. Using the average error as a measure of results, instead of the constant error as Judd had done, Spencer reports three of his four subjects making improvement.

One of the earliest studies of effect of incentives was made by Wright,¹⁹ on work and fatigue. Four subjects were required to work to exhaustion with the ergograph, but under three different conditions. In the first condition the subjects had no knowledge of how well or how much they were doing; they were merely told to work as hard and long as they could. The second condition furnished incentive by placing blocks under the carriage of the ergograph, and requesting the subjects to see how many times they could reach the block. In the third the subjects watched the records of their activities on a smoked drum, upon which there had previously been a line drawn, and were requested to do their best to reach the line as many times as they could. The records showed that under both conditions of incentive, the subjects performed more work, and were able to keep at it a longer time, than when they had no goal or incentive set for them.

A similar study was made by Arps.¹ Three subjects worked on the ergograph. During some of the work periods the subjects could see the tracings they were making on a smoked drum; and during other work periods the subjects had no knowledge of their results. Arps finds increased efficiency of the periods during which the results are known over those during which the results are unknown.

Johanson¹⁰ worked with reaction-times under three conditions: ordinary finger reaction to sound stimulus; same, but subject received an electric shock if he did not get his finger from the key quickly enough, i.e., if his reaction was slow; same, but subject was told the time of his previous reaction before receiving the stimulus for the next reaction. Three subjects were employed. Three hundred and fifty reactions were taken in each series for two of the subjects, and five hundred for the third subject. Both of the latter conditions effected shorter times, the second condition having a greater effect than the third. Johanson concludes that this is due to the fact that "The factors of positive and negative incentive caused the state of keener attention to be maintained."

Rexroad¹⁴ gave electric shock for punishment in continuous

multiple choice reactions to color. The main part of his experiment employed sixty subjects, thirty of whom received shock for inaccurate responses. Five colors were presented one at a time in random order, and the subject reacted to each by pressing one of five keys, according to a code. Both correct and incorrect responses were recorded; and when an incorrect response was made the subject was shocked by means of electrodes fastened to the hand. Each subject reacted for two and a half minutes with each of ten codes. The punished subjects were 15 per cent more accurate, or learned faster. The author divides the effect of the punishment into three sorts: disruptive, incentive, and instructive.

Incentives in the nature of verbal encouragement and discouragement were used on simple performances by Gates and Rissland.⁶ The tests used were the three-hole test for motor coordination and the color-naming test. Seventy-four Barnard College students were divided into three groups. After the initial test, one group was complimented, one was given adverse criticism, and to the third nothing was said. This was in each case followed immediately by a second test. In general, encouragement effected more improvement than discouragement, and discouragement more than mere repetition; but the differences were slight.

Thorndike¹⁷ had five subjects estimate in centimeters the length of fifty strips of paper when they had a 10-centimeter strip with which to compare them. This was followed by seven or eight training periods, during which the experimenter said "right" or "wrong" after removing each strip from the subject's view. Then came the final test period, like the first series, in which the experimenter said nothing. A comparison of the final with the first test period showed the average per cent of reduction in error to be 61. Seven subjects who had about the same amount of training, but were not told "right" or "wrong," made an average reduction in error of —7 per cent. In a second experiment reported in the same article, twenty-four subjects, blindfolded, drew a 3- (or 4-, 5-, or 6-) inch line in response to the appropriate command. The first test consisted in drawing one hundred and fifty lines of each of the four lengths. This was followed by seven training periods, like the first, except that the experimenter said "right" if the line was drawn within one-eighth inch of the correct length in the case of the 3-inch line and

one-fourth inch in the case of the other three lengths, and "wrong" if the line were longer or shorter. This was followed by a series like the first test without any comment by the experimenter. An average gain in per cent of correct lines, on the last test over the first, of 12 was made. Six subjects went through the training periods without any statement of "right" or "wrong" being made. Three of these improved and three made more errors.

Despite the facts that few subjects were used, that they did not all have the same number of training periods, that the intervals of time between training periods varied, and that the results for the six control subjects in the second experiment were not at all clear cut, and despite other possible interpretations, Thorndike considers "That these experiments are crucial as a demonstration that the consequences of a connection work back upon it to influence it."

3. STUDIES IN VISUAL DISCRIMINATION OF LENGTH

In view of the type of performance employed in this study, namely the visual discrimination of length on the Galton Bar, it is of interest to note the studies in which the performance was similar, even though they were made under normal conditions, i.e., without incentives. We shall here consider only those studies in which comparisons of horizontal lengths were made by the method of average error, and when the normal and comparison distances were presented simultaneously, for these are the more similar.

Fechner⁴ made the horizontal difference between two points of a compass equal to a standard distance set between the two points of another compass. Both compasses lay on the table side by side and only the points were visible. For each standard length used, he made 120 trials with the comparison length on the right of the standard and the same number with it on the left.

Volkman, reported by Fechner⁴ (p. 215ff.), used three vertical threads against a dark background. For each series of trials, the distance between two of the threads was constant, constituting the standard length, while the third thread had to be adjusted until the distance between it and the middle thread appeared equal to the standard. Volkman was the only subject and made 96 trials with each standard, half of them with it on the left and half on the right.

Appel, also reported by Fechner⁴ (p. 222), a student of Volkmann, extended the latter's work, using his apparatus and technique.

Chodin² used a fine black pencil line drawn on a paper. The standard length was marked off near the middle of this line by two short cross lines. Chodin indicated lengths equal to the standard by placing a short cross line both to the right and to the left of the standard. In one experiment 150 trials were made with each standard, and 120 trials in another.

Münsterberg¹³ had, for one of his experiments, distances marked off by white cardboard points visible against a green background. A distance of 60 mm. separated the standard from the comparison length. The latter was varied by sliding the farther point. Münsterberg as subject made 20 trials with each standard, half of which were with the comparison length on the left and half on the right.

Higier⁸ used two strips of black paper against a strip of glass with a $\frac{1}{2}$ mm. slit between them through which light came. A vertical thread across the slit constituted a dividing line, and black pieces slid in front of the bar to adjust the length on either side of the thread. The black pieces were moved by means of threads which were fastened to them. The subject's head was in a head rest which was 50 cm. from the apparatus. Only the right eye was used. Higier was the only subject and he made 50 trials, half right and half left, each day for ten days with each standard.

Fischer⁵ used the Münsterberg apparatus, except that the points marking off the length were of metal. He employed fewer standard lengths but more trials, namely eighty, for each.

Stephanowitsch¹⁶ employed a self-registering apparatus. The standard length was that between two marks placed on a paper beneath a glass plate. The glass plate could be made to slide over the paper by turning a wheel. A scratch on the glass had to be moved as far from one of the marks as was the other, thus making a distance between one of the marks and the scratch equal to that between the two marks. A marker attached to the glass recorded the length of its movement. In the major part of his experiments Stephanowitsch performed 100 trials for each of the standard lengths used.

Kiesow¹² used black lines $\frac{1}{3}$ mm. in width on two pieces of cardboard. A piece of paper covered the comparison length

and was drawn along until the subject said to stop, after which he was allowed to make a finer adjustment himself. The cardboards lay horizontally on a table before the subject. The one subject performed 100 trials with each standard length.

All these experiments on the visual discrimination of length were performed with one of two purposes, namely, to test the applicability of Weber's law or to study psychophysical methods, which do not concern us here. To these ends each experimenter employed a number of different standard lengths. With the exception of Stephanowitsch¹⁶ who used five subjects, the results are those of but one subject, usually the experimenter himself, and are generalized without hesitation.

One of the lengths employed by six of the above investigators is of more interest to us than the rest of their work because of its equivalency with that used in some of the work about to be reported. Although the results of these investigations are not directly comparable because of the great differences in the apparatus employed, in the number of trials, in the distance of the subjects from the lengths being compared, and in the freedom of movement of the head and eyes, the findings when the standard length was 200 mm. are shown together in Table 1.

TABLE 1

<i>Experimenter</i>	<i>No. of Subjects</i>	<i>Distance from Eyes mm.</i>	<i>No. of Trials</i>	<i>Measure</i>	<i>Standard 200 mm. %</i>
Volkman	1	800	96	Rel. A.E.	0.96
Appel	1	370	192	Rel. A.E.	3.09
Appel	1	300	132	Rel. A.E.	3.38
Münsterberg	1	600	20	Rel. V.E.	2.50
Higier	1	500	500	Rel. A.E.	2.55
Kiesow	1	400*	100	Rel. A.E.	0.65

* Approximate.

Of more value to the present experiment would be data showing the effect of practice. Unfortunately these investigators do not report their original data but present only very meagre summary tables. The only one who presents any data to show the course of the average error throughout the experiment is the most recent investigator, Kiesow.¹² His data on this point will be discussed on p. 42.

We find that in animal work incentives in the nature of "reward" and of "punishment" have been used and compared, when applied to brightness discrimination. In general, a combination of the two seems to have been most effective, with punishment ranking next, and reward last. In human studies we find incentives effecting improvement in such simple activities as ergographic work and speed of finger reactions.

The first exact knowledge of the operation of any factor or process must come through observation of a simple delimited case of that factor acting under known and constant conditions upon a simple, isolated, measurable bit of behavior. To deal immediately with complex cases, containing many variables, results only in confusion.

A type of performance which meets these demands, but upon which, our survey reveals, the effect of incentives* has scarcely been studied, is that of visual discrimination. This experiment attempts to employ simple incentive stimuli, whose natures and intensities are constant and comparable; and to use them under constant and known conditions; and to have them act upon this simple and measurable bit of behavior, visual discrimination of length. These stimuli take on a positive or negative character according to the nature of the instructions issued. Thus, when they occur as a consequence of correct reactions they may be said to be positive (reward), and when they follow incorrect reactions they may be said to be negative (punishment).

It is not surprising to find improvement in any performance following upon receipt of an incentive stimulus, when it includes some indication of the type or nature of the error which has been made, that is, some indication of the direction or steps to be taken in effecting improvement. But it is not so clear that improvement will follow when the incentive stimulus consists only in something like punishment or reward, without any indication of the nature of the error committed, or of the steps to be taken to avoid it. Visual discrimination of length is a performance which lends itself well to an investigation of this point, and the major portion of the present experiment has been arranged to that end. It is

* Throughout this report the terms incentive, reward, and punishment are used to indicate the interpretation or attitude of the subject toward the situation, while incentive stimulus refers to the external physical stimulus.

possible to use an incentive stimulus which satisfies the demands stated in the paragraphs above, and which may be interpreted as a reward for good discrimination or as a punishment for poor discrimination, without giving any indication of the direction of the error.

II. EXPERIMENTAL SITUATION

1. APPARATUS

The apparatus was designed to meet three chief requirements: (a) to measure visual discrimination of length without having any visual stimuli present except the actual length being judged; (b) to present incentive stimuli which would take on the nature of punishment for poor discrimination and of reward for good discrimination, without giving any indication of the nature of the error, and which would be comparable in respect to quality and intensity, and to present them automatically; (c) to measure the time taken for each discrimination made, without the subject having knowledge of being timed. These requirements were met in the following manner.

(a) A modification of the Galton Bar was employed in presenting the lengths to be discriminated, and in measuring the accuracy of the performance. An illuminated glass bar was mounted horizontally in the center of a black field, in a dark room. The illuminated bar was three-quarters of an inch wide and thirty-six inches long. The black field was four feet wide in the vertical direction, and eight feet in the horizontal direction. The lower edge of the bar was four feet six inches from the floor.

The bar itself was composed of opal glass. The source of illumination was a row of ten 15-W. lamps, whose centers were three and three-quarters inches apart, of the same length as the bar, and running parallel behind and above it. The light was then reflected onto the bar by a mirror, placed at an angle of 45° , as shown in Diagram 1. This combination results in a strip or length of light of sensibly even intensity of illumination.

A vertical hair line marked the center of the bar. Two black shields were made to slide close against the bar, one on either side of the hair line, thus cutting out the light, and making it possible to vary the length of the strip of light seen on either side of the center hair line. The shield on the subject's left of the center could be set by the experimenter at any point, making a standard line of light of any desired length. The subject could vary the length to his right of the center, by turning a control rod, as used with the regular Galton Bar,

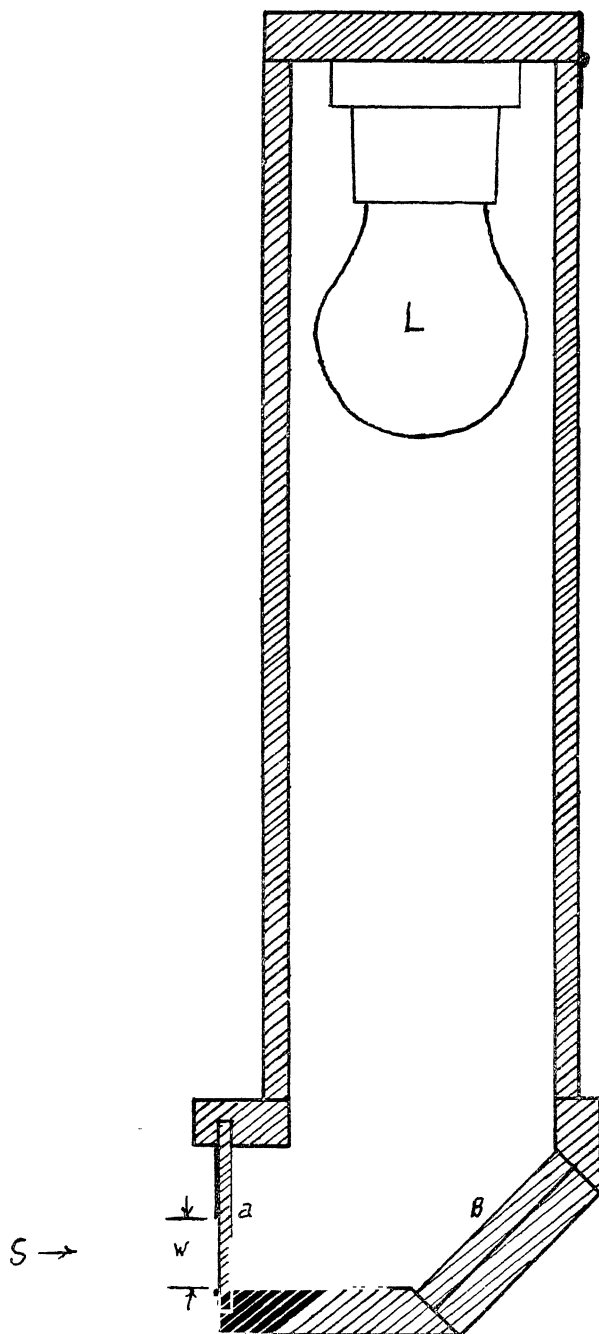


DIAGRAM I.

SKETCH SHOWING CROSS-SECTION OF BAR, AND METHOD OF LIGHTING IT.

L = lamp a = opal glass.

B = mirror w = part of "a" through which light is allowed to pass.

as made by C. H. Stoelting & Company. A scale back of the bar made it possible to read in millimeters the distance or length exposed on each side of the center.

It will be seen that this is in effect a Galton Bar; but one that consists in lengths of light seen in the dark, thus eliminating all visual stimuli except the actual lengths being discriminated.

The subject was seated with his eyes approximately on a level with the bar, and six feet in front of the center of it.

(b) The incentive stimulus, which took on the nature of punishment for poor discrimination, or of reward for good discrimination, consisted in the sound of an electric door-bell. The stimulus was presented in the following manner. On a table immediately in front of the subject was a push button. After each discrimination the subject pressed this push button, which lighted a small, blackened lamp on the experimenter's side of the apparatus, indicating that the subject had completed his discrimination, or was through judging. Now the bell was also in circuit with this push button. To give punishment, the bell rang when the subject pressed his button, if he had made the length too long or too short, but did not ring if the correct length had been made. To give reward, the bell rang when the subject pressed his button, if he had made the correct length, but did not ring if the length had been made too long or too short. In circuit with the bell was a mechanism which automatically closed the circuit, whenever the length was too long or too short, in the case of punishment, so that the bell would respond according to the above formula, whenever the subject pressed his button. If it were reward that was being given, the same mechanism automatically opened the circuit when the length was too long or too short, and closed it when the length was correct, thus here also properly controlling the ringing of the bell when the subject pressed his button.

The mechanism, which determined whether or not the bell would ring when the subject pressed his button, was composed as follows: A long brass rod was attached to the shield which the subject moved in adjusting the length, and moved with and parallel to it. Two contact points or fingers extended out from the rod, and by means of collars and set screws each could be set at any point along the rod. As the rod moved, these contact points or fingers slid along a surface

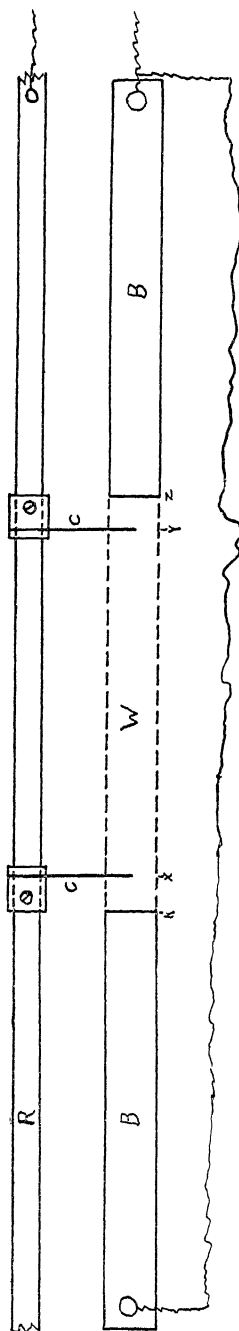


DIAGRAM 2.

SKETCH OF MECHANISM CONTROLLING THE RINGING OF THE BELL
 R = brass rod. C, C = adjustable contact points. B, B = brass strips.
 W = strip of wood. Wavy line = wiring

consisting of two brass strips separated by a strip of wood, as shown in Diagram 2. Now in the case of punishment, any desired amount of variation from the true length might be allowed before the bell would ring, by setting the contact points that distance along the wood strip from the brass strips. As shown in the diagram, the contact points are as they might be when the correct length of light has been set, i.e., equidistant from the brass strips. But if the subject had set the length as much as the distance kx in error in either direction, the bell would ring. Thus the subject was allowed an error of kx without causing the bell to ring when he pushed his button.

In the case of reward, reversing a double pole, double throw switch, as shown in Diagram 3, cut this mechanism out, putting the bell in direct circuit with the subject's push button. At the same time the mechanism was placed in circuit with a relay, which could break the bell circuit. Thus, if the contact points were on the wood strip—i.e., if the length was correct—the relay did not operate, and the bell would ring when the subject pressed his button. But if one of the contact points was on a brass strip—i.e., if the length was either too long or too short—the relay broke the bell circuit, and the bell would not ring when the subject closed his push button.

The determination of what distance was to be set as too long and too short, i.e., what amount of variation from the true length would be called wrong, will be explained later with the procedure.

This method of presenting the incentive stimuli automatically, gives mechanical accuracy and certainty as to the conditions and manner of presentation, and so rules out the chances of error and variability which might be great were the experimenter himself presenting or controlling the incentive stimuli. At the same time it permits of the greatest flexibility in the amount of error which is to be considered as error, requiring only that this be determined, as it should be, before and not during the making of a trial.

It may also be pointed out that the requirement of equating the incentive stimuli in respect to quality and intensity had been met, for they consisted in the same bell operated by the same strength of current. Yet in the one case, the bell very definitely took on the nature of punishment, and in the other case the nature of reward.

(c) The time taken in making each discrimination was measured by a Stoelting electrically controlled stop watch. A contact point was fastened to, but insulated from, the brass rod, described above, which moved with the shield regulating the length of the line of light. This point was so placed that it passed over a copper point on a stationary part of the apparatus, as soon as the subject moved the shield away from the zero point or center, when starting to adjust the bar of light. The contact thus made started the watch. The subject's push button also operated a relay which closed the watch circuit,

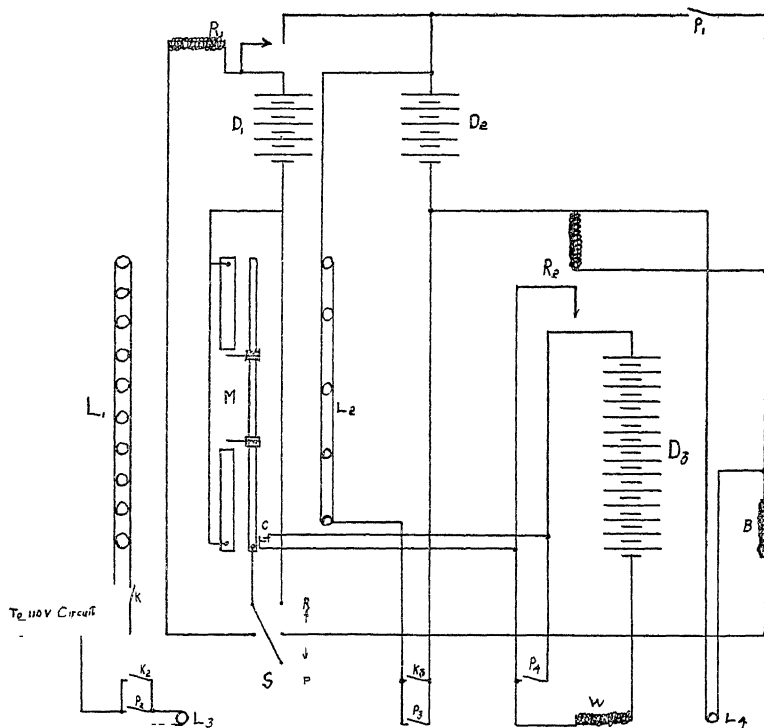


DIAGRAM 3

WIRING DIAGRAM

L_1 , ten 15W. lamps lighting the bar. L_2 , five 6V. lamps to illuminate the mm. scale. L_3 , lamp to illuminate the watch. L_4 , indicator lamp operated by subject's push button. P_1 , subject's push button. P_2 , P_3 , P_4 , push buttons. K_1 , K_2 , K_3 , knife switches. S , double pole, double throw switch; when closed in the direction P , the circuits are set for punishment; when closed in the direction R , the circuits are set for reward. M , mechanism controlling the ringing of the bell. B , bell. R_1 , relay for breaking the bell circuit. R_2 , relay for stopping watch. C , contact points for automatically starting watch. W , watch. D_1 , D_2 , D_3 , dry cells.

stopping the watch. Thus the time was recorded from the point when the subject began to make his line until he pressed the button, indicating that he had completed his judgment. Then as the experimenter moved the shield back to the zero point or center, the contact points which started the watch again closed the circuit in passing, automatically bringing the watch back to the zero point. A third switch was placed in the watch circuit, in parallel with the other two, and was located on the experimenter's control board. This was for the purpose of readjusting the action of the watch to the subject's procedure if, as occasionally happened, it was thrown out of step by a subject pressing his push button twice, instead of once.

The noise of the watch and the magnet operating it was ruled out by placing this instrument on a thick pad of felt and covering it with two bell jars. This arrangement was found to be effective, and still allowed the experimenter to read the watch easily, and to remove it to do the necessary winding. The subject had no occasion to suspect that he was being timed.

Both relays employed, the one in one of the bell circuits and the other in the watch circuit, were sound-proofed in a manner similar to that used with the watch.

The electric circuits whose uses are described above are all shown in Diagram 3.

2. SUBJECTS

Sixty undergraduate students in psychology in Barnard College, Columbia University, comprised the subjects for this experiment. These were, of course, all girls. The six groups described below contained ten subjects each.

3. PROCEDURE

The task was the same for all of the groups. The incentive was the variable. The experimenter exposed a standard length of 120 millimeters on the subject's left of the center line, and the subject was required to make the length on his right *twice as long as the standard*. Each trial started with the shield at the center line, so that there was no length showing on the right of the center. By turning the flexible control rod, the subject moved the shield out until he thought he had exposed a line twice as long as the standard. He was not allowed to

move the shield back toward the center. Nothing was said about time, the subject being allowed to fall into his own pace in setting the lengths. When satisfied that he had exposed enough line, the subject pressed his push button. This stopped the watch, and lighted the experimenter's indicator lamp. The experimenter then observed and recorded the length set and the time taken; and moved the shield back to the center. The subject immediately proceeded with the next trial.

The experiment occupied about an hour on two different days. On the first day all of the groups were given a practice series of fifty trials, in ten series of five trials each, with a rest period of approximately one minute between each series of five. On the second day, at the same time of day, the subject first performed a series of five trials, in the same manner as on the first day. At this point the incentive series was introduced. This consisted in forty-five trials, under the condition of incentive, and was performed in nine series of five trials each, with a rest period of approximately one minute between each series of five. As the incentive was varied from group to group, this series must be described separately for each group.

a. Punishment Group

At the beginning of the incentive series the subjects in the Punishment Group were told that from this point on the procedure would be the same as before with this difference: that if they got the length wrong, a bell would ring when they pushed the button. They were also told that of course one would rarely ever get the length *exactly* right, so that they would be allowed a little leeway before it would be considered wrong.

The criterion of wrong, i.e., the amount of variation in either direction from the correct length which would be allowed without causing the bell to ring, was determined as follows. The average error each subject made on his own fifty practice trials was taken as a basis for him. Before the series began, the sliding contact points, described above, were set to allow that amount of variation from the correct length, before the bell would ring. For example, if a subject made an average error of twelve millimeters on his fifty practice trials, the mechanism was set so that the bell would ring only when he had made the length as much as twelve millimeters

too short, or as much as twelve millimeters too long. But, as the experiment progressed, whenever a subject's average error for any series of five trials became smaller than his average error had previously been, the amount of variation allowed was narrowed to equal the new and smaller average error. The rest period between each series of five trials permitted the experimenter to calculate the average error for the series just completed, and to make the necessary changes, if any, in the mechanism which controlled the ringing of the bell. The subject, of course, had no knowledge of this changing criterion of error.

This method helps to equate the punishment from subject to subject, since the criterion of error, and therefore the administration of the punishment, was always dependent upon each subject's own previous performance.

b. Reward Group

The procedure for the Reward Group was the same in all respects as that for the Punishment Group, with this important difference: that the bell rang when the length had been made correctly, while nothing occurred when the length had been made too short or too long, and the subject was so instructed.* The criterion of right and wrong, and hence the administration of the incentive stimuli, was adjusted as above to each subject's own performance as the experiment progressed.

* Clearly the silence which sometimes followed the pressing of the push button was a part of the total stimulus situation which confronted the subject. Thus to the subjects in the Punishment Group the bell indicated failure and silence indicated success, and in the case of those in the Reward Group the bell indicated success and silence failure. Bell and silence might be thought of as two parts of a total configuration which confronted the subject. But it was the bell which was the differentiating, determining feature of the stimulus situation. The subjects of the Punishment Group, for instance, were definitely trying to avoid the bell, they were not trying to attain silence. Their expressive responses, as swearing and chagrin, were made when the bell rang, they did not show delight at the silence. Likewise the Reward Group were oriented toward the bell, not the silence, and gave expressions of delight when the bell rang, but were not expressive in response to silence. Also, one is confronted with a metaphysical problem when he attempts to differentiate the silence which followed the pressing of the push button from that which had been existing all of the time and which continued to exist until the next time the bell rang. In view of these considerations the bell has been referred to exclusively as the incentive stimulus, throughout this paper; but perhaps one should bear in mind the presence of silence throughout the experiment as a kind of neutral reciprocal factor.

c. Control Group

The subjects in this group performed the entire fifty trials on the second day in exactly the same manner as they had in the practice series, the first day; but the bell was not introduced.

d. Guess-with-Punishment Group

The procedure for the Guess-with-Punishment Group was exactly the same as for the Punishment Group, with this addition: that each time the bell rang, i.e., each time an error had been made, the subject was required to guess the direction of his error.

e. Told-with-Punishment Group

For this group the procedure was exactly the same as for the Punishment Group, with this addition: that each time the bell rang the experimenter told the subject the direction of his error. The experimenter did this by simply saying either "long" or "short," after the bell had rung.

f. Knowledge Group

The determination of an error here was the same as for all the other incentive groups. That is, as the average error decreased, the amount of variation allowed was narrowed accordingly. But the bell was not used. Instead, after each trial, the experimenter said simply either "long," or "short," or "right." These words were always said, as nearly as possible, in the same matter-of-fact tone of voice.

III. STATEMENT OF RESULTS

The sort of measure which has to be dealt with is amount of error in millimeters. Also, since the unit of performance in the procedure of the experiment was a series of five trials, the most appropriate grouping of the data is by the same series of five trials. Hence the average error of a series of five trials has been employed as the unit in the consideration and presentation of the data. The average error in millimeters, of each series of five trials made by each subject is given in the Appendix, pages 49-54. The results for the Control Group may be found in Table 12; the Punishment Group, Table 13; the Reward Group, Table 14; the Guess-with-Punishment Group in Table 15; the Told-with-Punishment Group in Table 16; and the Knowledge Group, Table 17.

The effect of the incentive in the experiment is measured by the change in the magnitude of the average errors made during the incentive series from those made during the practice series, when this change is considered in relation to the performance of the Control Group. To obtain an adequate index of a group of subjects, the amount of change made by each subject must first be considered in relation to that subject's own performance during the practice series. Otherwise, the variation from subject to subject in absolute amount of error would vitiate the group index. Therefore, the average error made by each subject for each series of five trials has been considered as a percentage of his average error in his own fifty practice trials. Thus it becomes possible to average the performance of all of the subjects within a group for each series of five trials, obtaining an index of group performance throughout the progress of the experiment.

This method of making the group measure of change, effected by the incentive stimulus, an average of the relative and not the absolute change in each individual's performance is at the same time perhaps the best method of equating the groups, in respect to ability in this type of performance.

The average of each group shown in the above tables, i.e., the group index of the percentage which the average error of each series of five trials is of the average error of the fifty practice trials, is given in Table 2, page 25.

Analysis of the table reveals that all of the groups per-

TABLE 2

AVERAGE ERROR OF EACH FIVE TRIALS AS A PERCENTAGE OF THE AVERAGE OF THE FIFTY PRACTICE TRIALS.

<i>Group</i>		<i>C</i>	<i>P</i>	<i>R</i>	<i>GwP</i>	<i>TwP</i>	<i>K</i>
Series of five trials							
P R A C T I C E	1st	134	165	146	139	191	108
	2nd	108	98	99	97	108	98
	3rd	84	82	96	105	81	96
	4th	103	98	92	84	95	85
	5th	82	73	80	94	83	103
	6th	87	104	73	95	97	111
	7th	98	90	94	109	107	77
	8th	96	102	109	95	71	109
	9th	104	89	104	100	75	93
	10th	109	99	108	84	90	118
I N C E N T I V E	1st	102	107	79	75	102	98
	2nd	84	85	97	70	76	71
	3rd	92	82	72	66	82	71
	4th	97	71	79	70	58	77
	5th	113	60	67	69	65	57
	6th	117	63	76	46	38	57
	7th	105	47	69	49	43	38
	8th	101	38	55	46	38	56
	9th	116	42	37	39	31	53
	10th	127	24	26	15	20	45

formed in the same manner on the practice series, i.e., the fifty trials of the first day. The average error of the second five trials shows a considerable decrease over that of the first five trials; and that of the third five shows a further decrease. But from that point on there is no further decrease, the curves fluctuating around a practically level line, a sort of normal error. The Knowledge Group is an exception only in that it was not so far from its normal error at the beginning.

The incentive series, i.e., the second day's trials, under the condition of incentive, present quite a different picture. In the first place, the Control Group, which had no incentive stimulus, merely continuing its practice, and therefore forming a standard or norm with which to compare the different incentive groups, made but little change in its errors. And this change was in the nature of a slight increase, the average errors through successive series of five trials starting at 102% and increasing to 127% of the average error of its fifty practice trials.

The Punishment Group started at 107% and decreased its

average error in a very regular manner to only 24% of the average error made on its fifty practice trials.

The Reward Group started at 79% and decreased its average error to 26% of its first day's performance.

The Guess-with-Punishment Group decreased its average error from 75% to only 15%; the Told-with-Punishment Group from 102% to 20%; and the Knowledge Group from 98% to 45%.

As all the conditions of incentive employed effected a decrease in the average error, and not even a single subject within any group has to be excepted from this statement, the first question of the experiment is thus answered: incentives did effect a gain in accuracy of discrimination.

Therefore, the second question of the experiment resolves itself into the following. Throughout the incentive series, i.e., during the period of work measured by the experiment, how much can the error be decreased, how much improvement can be effected, by each condition of incentive employed? The best measure of this would be the performance on the last unit of the series, i.e., the average of the last five trials. This is shown in Table 3.

TABLE 3.
AVERAGE ERROR OF THE LAST FIVE TRIALS OF THE INCENTIVE SERIES AS A PERCENTAGE OF THE AVERAGE OF THE FIFTY PRACTICE TRIALS.

<i>Group</i>	<i>C</i>	<i>P</i>	<i>R</i>	<i>GwP</i>	<i>TwP</i>	<i>K</i>
Percentage	127	24	26	15	20	45

To give meaning to the differences between the groups, their reliability has been computed by the standard method, and is shown in Table 3, page 27.

Table 4 shows: (1) that compared with the Control Group each of the conditions of incentive may be said to have effected a reliable decrease in error, or improvement in discrimination. This difference is not only statistically reliable, but is large. (2) That there is no reliable difference between the performance of the Punishment and of the Reward Groups. (3) That there is no difference between the Guess-with-Punishment and the Told-with-Punishment Groups; nor between these groups and the Punishment and Reward Groups. (4) That the Knowledge Group made a reliably less amount of

TABLE 4.

RELIABILITY OF THE DIFFERENCES BETWEEN THE EFFECTS OF THE DIFFERENT INCENTIVES, AS MEASURED BY THE AVERAGE ERROR OF THE LAST FIVE TRIALS AS A PERCENTAGE OF THE AVERAGE OF THE FIFTY PRACTICE TRIALS.

<i>Groups</i>	<i>Diff.</i>	<i>Sigma Diff.</i>	<i>Diff.</i> <i>Sigma Diff.</i>
C and P	103	17.1	6.0
C and R	101	17.3	5.8
C and GwP	112	16.8	6.7
C and TwP	107	16.9	6.3
C and K	82	17.8	4.6
P and R	2	5.9	0.3
P and GwP	9	4.1	2.2
P and TwP	4	4.7	0.9
P and K	21	7.2	2.9
R and GwP	11	4.7	2.3
R and TwP	6	5.3	1.1
R and K	19	7.7	2.5
GwP and TwP	5	3.2	1.6
GwP and K	30	6.4	4.7
TwP and K	25	6.8	3.7

improvement than did the Guess-with-Punishment and the Told-with Punishment Groups; and that near reliability was reached between the Knowledge Group and the Punishment and Reward Groups; at the same time, however, the Knowledge Group showed real improvement in its performance, thus occupying a sort of mid-point.

IV. DISCUSSION OF RESULTS

1. NO IMPROVEMENT THROUGH REPETITION ALONE

The performance of all the groups during their practice series, and especially the Control Group throughout its entire performance, shows that, beyond the first few trials, the mere repetition of this act of discrimination did not effect any improvement. Practice alone, without any incentive or any knowledge of the quality of the performance, did not bring about any decrease in the error. This is in agreement with the findings of Judd,¹¹ who also devised a situation in which the subject received no knowledge of the accuracy of his performance, and found that he made no improvement. It is probably rarely true that an individual is in a situation where there is neither incentive nor knowledge of the quality of the performance. But here, in those parts of the experiment under discussion, there was certainly no knowledge of the quality of the performance. And if there was any incentive, it was of the general sort, as a general desire to do well, or to please the experimenter, which proved to be not in the least effective, unless the improvement of the first five trials might be ascribed to it.

2. EQUIVALENCY OF THE EFFECTS OF PUNISHMENT AND OF REWARD

The effect of the incentive stimulus, when it took on the nature of punishment, and when it took on the nature of reward, was found to be the same. Under incentive, the Punishment Group decreased its average error to 24% of the average error of its practice series, and the Reward Group decreased its average error to 26% of the average error of its practice series. These may be considered as equivalent decreases.

In previous studies, where a comparison has been attempted between these two contrasting types of incentives, the usual finding has been that of the greater effect of punishment. But it should be noted that in all of these studies there has been considerable discrepancy in the nature of the incentive stimuli compared, their quality, their intensity, the manner of presentation, etc. For example, in animal studies, shock has been the standard punishment stimulus, and food the standard reward stimulus. And in studies with human subjects, perhaps shock was the punishment stimulus, and a knowledge of scores

made, the reward stimulus. To attempt to compare the effectiveness of shock and of food is like trying to compare two loaves of bread by saying that one of them was 12 inches long, and the other weighed 12 ounces. They have entirely different qualities, and so cannot be adequately equated, or even compared.

In this study we have carefully controlled the quality, intensity, and method of presentation of the stimulus for these two conditions of incentive. The *physical* aspects of the stimuli were equivalent, for they were identical; and they were presented in the same carefully regulated manner, as described with the apparatus and the procedure. When the incentive stimulus took on the nature of punishment, the bell rang for failures or errors in discrimination. The incidental, spontaneous behavior of the subjects in the Punishment Group gave evidence of the fact that a real punishment was being given. The subjects often swore at the bell, made "faces" when it would ring, and said that it was quite a shock to them, and all adopted the attitude of trying to keep the bell from ringing. While these spontaneous bits of behavior on the part of the subjects in the Reward Group, when the bell rang for successes, were quite the opposite. They would give exclamations of delight upon hearing the bell, and sigh, and say, "What a relief," "That's better," etc. And all tried hard to make the bell ring.

Whether an incentive stimulus is in the nature of a punishment or of a reward depends not upon the physical stimulus itself, but upon the conditions under which it is given, and so might be said to be an interpretation on the part of the subject, a subjective thing. It is readily conceivable that under certain conditions an electric shock might be considered as a reward, and that food might become punishment. The factors, the effects of which we are trying to measure, then, are not the physical stimuli at all, but the additional elements given to them by the conditions under which they are presented, and by the subject who responds to them. Obviously, the physical stimuli themselves, and the conditions under which they are given, must be equated, however, before we can begin to make comparisons. For if they are different, we may then be measuring those differences, the effects of those factors, rather than the effects of the punishment and reward themselves.

The results of this study, then, force us to conclude that when the physical factors are equated, and when the punishment and reward are placed over a common denominator, as they have been here, they too are equivalent in effect. Or stated another way, as far as the actual effect upon the subject's visual discrimination of length as measured on the Galton Bar is concerned, it makes no difference whether the incentive stimulus is one to be sought after or to be avoided, reward or punishment.

3. ANALYSIS OF THE IMPROVEMENT

a. The Constant Error and the Variable Error

Examination of the constant error and of the variable error throws some interesting light on the improvement made.

Table 5 shows the magnitude of the constant error in millimeters for each of the groups. The magnitude of the constant error is given for successive groups of 15 trials. The first item in the table is the constant error for the 2nd, 3rd, and 4th series of five trials; the second item in the table is the constant error for the 5th, 6th, and 7th series of five trials; etc. Some of the subjects within a group had constant errors carrying a plus sign, and others had constant errors carry-

TABLE 5.
MAGNITUDE OF THE CONSTANT ERROR IN MM.

Group				C	P	R	GwP	TwP	K
Series of five trials									
P	2nd, 3rd,	4th		13.8	16.6	10.7	16.5	8.0	12.2
R									
A	5th, 6th,	7th		12.3	16.2	6.3	19.5	8.8	16.5
C									
T	8th, 9th,	10th		11.0	19.5	9.7	16.4	6.9	16.1
I									
C	2nd, 3rd,	4th		11.5	8.9	6.1	7.5	3.4	6.2
E									
N	5th, 6th,	7th		15.5	4.4	5.5	5.4	2.3	3.2
T									
I	8th, 9th,	10th		12.8	2.4	2.1	3.3	1.9	1.6
V									
E									

The constant errors of the individual subjects, from which are computed the group values, shown in Table 5, are given in Tables 18-23, pages 55-60.

ing a minus sign but the constant errors of the individual subjects within a group have here been averaged without regard to signs. Hence, the table shows relationships of only the magnitude of the constant errors.

We see practically no change in the magnitude of the constant error throughout the practice series of all the groups, and the Control Group continues to show no change throughout its second day's trials. In each of the five incentive groups, however, the magnitude of the constant error shows regular and marked decrease, throughout the second day. This is of importance in view of the fact that, except in the TwP and K groups, the incentive stimulus gave to the subject no indication of the direction of his error.

Table 6 shows the variable error in millimeters for each of the groups. The table follows the same form as the one for constant error. The variable error is given for successive groups of 15 trials, and these are designated as series of five trials.

As in the case of the constant error, we see practically no change in the variable error throughout the practice series of all the groups, nor any change throughout the second day's trials of the Control Group. But the incentive groups show

TABLE 6.
VARIABLE ERROR—IN MM.

Group			<i>C</i>	<i>P</i>	<i>R</i>	<i>GwP</i>	<i>TwP</i>	<i>K</i>
Series of five trials								
P	2nd, 3rd,	4th	8.2	8.3	7.6	8.4	8.0	8.5
R								
A	5th, 6th,	7th	8.0	8.8	7.4	6.8	8.6	7.5
C								
T	8th, 9th,	10th	8.9	8.6	7.7	7.2	7.6	8.3
I								
C	2nd, 3rd,	4th	7.3	8.2	7.8	7.4	7.2	8.1
E								
N	5th, 6th,	7th	7.4	6.8	6.4	7.3	5.8	6.4
T								
I	8th, 9th,	10th	8.4	4.6	4.3	4.7	3.5	6.6
V								
E								

The variable errors of the individual subjects, from which are computed the group values, shown in Table 6, are given in Tables 24-29, pages 61-66.

a regular and marked decrease in their variable errors throughout the second day. The Knowledge Group is the single exception to this, its variable error showing but little decrease.

A possible hypothesis which suggests itself in explanation of the improvement made under incentive is that the subject upon realizing that he had made an error was able to judge its direction, and so modified his next response in the right direction. Such an hypothesis would account for the decrease in the constant error, but it is not adequate to account for the decrease in the variable error. For the decrease in the variable error shows that the improvement made under the various conditions of incentive was not merely one of orientation in respect to direction of error, but also brought with it decrease in variability of performance.

However, in one case, that of the Knowledge Group, the decrease in constant error was not accompanied by a decrease in variable error. Thus, here, in the relationships between the constant errors and the variable errors, is to be found substantiation of the finding that the knowledge situation was not as effective as the others.

b. Time of Making Discrimination

The time taken in making a discrimination might be of value when considered in relation to the accuracy of the performance.

The records of the time taken in making each discrimination have been treated in the same manner as the records of error. The average time per trial for each series of five trials has been taken as the unit. To make meaningful the average of a group of subjects, the average time for each series of five trials for each subject has been considered as a percentage of the average time per trial of that subject's own 50 practice trials. And the group average for each series of five trials consists in the average of these percentages.

Tables 30-35, pages 67-72, show the absolute time per trial in seconds for each series of five trials, for each subject.

Table 7, below, then, shows the group averages of the average time of each five trials as a percentage of the average of the 50 practice trials.

We note that: (1) during the practice series all of the groups showed a slight, gradual decrease in time; (2) during

the second day's trials the Control, Punishment, and Reward Groups continued their slight, gradual decrease in time, in a practically equivalent manner; (3) as soon as the incentive stimuli were introduced, the time of the Guess-with-Punishment Group made a sharp increase, from which point there ensued again the slow, gradual decrease; but the time did not finally become as low as it had during the practice trials; (4) upon the introduction of the incentive stimuli, the Told-with-Punishment Group began a slight, gradual increase in its time; (5) the Knowledge Group, upon the introduction of the incentive stimuli, made a small, immediate increase, and thereafter remained on a level.

The improvement in discrimination in the Punishment and Reward Groups did not interrupt the gradual decrease in the

TABLE 7.

AVERAGE TIME OF EACH FIVE TRIALS AS A PERCENTAGE OF THE AVERAGE OF THE FIFTY PRACTICE TRIALS.

Group		<i>C</i>	<i>P</i>	<i>R</i>	<i>GwP</i>	<i>TwP</i>	<i>K</i>
Series of five trials							
	1st	118	124	133	128	102	116
P	2nd	112	105	114	117	105	113
R	3rd	101	115	111	114	103	107
A	4th	96	97	108	109	106	99
C	5th	101	98	104	99	95	100
T	6th	90	92	86	85	98	91
I	7th	95	101	89	94	96	95
C	8th	95	93	90	91	96	94
E	9th	93	95	91	85	95	91
	10th	84	88	86	81	92	94
I	1st	95	89	92	84	97	89
N	2nd	85	93	82	108	96	100
C	3rd	90	93	80	106	104	95
E	4th	93	92	76	109	100	93
N	5th	93	94	72	102	105	93
T	6th	88	89	70	97	106	97
I	7th	88	87	72	98	108	93
V	8th	81	80	74	91	109	98
E	9th	73	81	83	91	104	96
	10th	75	76	69	87	106	93

time per trial throughout the experiment, which decrease went on exactly as it had in the Control Group, for which no incentive stimulus was introduced and no improvement made in discrimination.

The Guess-with-Punishment and the Told-with-Punishment Groups, in both of which some factors, in the nature of knowledge of direction of error, were introduced in addition to pun-

ishment, made slightly greater improvement in discrimination than did the pure Punishment and Reward Groups. These two groups were also the ones in which the presumably normal tendency to gradual decrease of the time per trial was interrupted. Besides receiving punishment, it was made certain that the subjects in these groups would pay particular attention to the direction of their errors. In the Guess-with-Punishment Group, the introduction of the additional factor slowed the time considerably; but the gradual decrease set in again. But in the Told-with-Punishment Group, the effect on the time was to slow it gradually.

Directing the subject's attention to the direction of his errors seemed to cause him to take more time, but brought scarcely any greater improvement in discrimination, and that only when used in addition to punishment, and not when used alone, as in the case of the Knowledge Group.

In view of the above facts and discussion, there is but one safe conclusion: that the time of making a discrimination does not bear any consistent relation to the accuracy of the discrimination; or that some factors which make for improvement in accuracy tend to increase the time, while others do not. Any attempt to read into the data specific relationships between improvement and time is purely speculative. Thus another possible hypothesis to explain the effect of incentives is ruled out.

c. Guess-with-Punishment Group

When considering the improvement made in discrimination, it is natural to wonder if the subjects could, and if they did, guess the direction of their errors, upon the receipt of punishment, or reward, i.e., when the bell rang. The Guess-with-Punishment Group was introduced in the hope of throwing light on this question. It will be recalled that the procedure with this group was identical with that for the Punishment Group, the bell ringing after each error, except for the added requirement that each time the bell rang the subject must guess the direction of his error.

Out of the total number of guesses as to direction of error, the group averaged 70% correct guesses, with the range of individual subjects running from 57% to 89%.

With the right guesses amounting to 70% of the total number of guesses, we note that the subjects were able to do better

at guessing the direction of their errors than chance factors alone would warrant, but that they were far from being able to tell with certainty.

If knowing the direction of the error is one of the chief cues through which the improvement is made, it might be expected that even 70% accuracy in guessing would be sufficient to account for the improvement; but there are two indicators in another direction: (1) a comparison of accuracy in guessing with decrease in average error by individuals; (2) the Told-with-Punishment Group.

If the individuals in the Guess-with-Punishment Group are ranked according to amount of improvement made in discrimination, and ranked again according to their accuracy in guessing the direction of their errors, the rank difference correlation is $-.18$. Such a coefficient indicates that those individuals making the best guesses as to the direction of their errors did not tend to make the most improvement in discrimination.

d. Told-with-Punishment Group

Another answer to this question was sought with the Told-with-Punishment Group. If knowing the direction of the error is one of the chief cues with the aid of which the improvement is made, it might be expected that greater improvement would result if the subject always knew the direction of his error. Hence, this group was given exactly the same conditions as the Punishment Group, but in addition the experimenter said "long" or "short" each time after the bell had rung, thus insuring that the subject knew the direction of his error.

It has been noted that this group reduced its average error to 20% of the average error made on its practice series, and that this is only slightly, and not at all reliably, better than the performance of the group which received punishment alone. And neither can the improvement effected in this group be considered as different from that effected in the Guess-with-Punishment Group.

Since the Guess-with-Punishment Group shows that subjects were able to guess the direction of their errors only 70% of the time, it would be expected that complete knowledge as to the direction of the error would bring about greater improvement in discrimination, providing that this is one of

the chief cues employed in decreasing the average error. But since the improvement of the Told-with-Punishment Group was not greater than that of either the Guess-with-Punishment Group or the Punishment Group, we can but conclude against the likelihood that guessing the direction of the error was an important cue in the bringing about of the improvement.

e. Knowledge Group

The Knowledge Group brings out some interesting relationships. The incentive or second day's series for this group was as follows. After each trial the experimenter said, simply "long," "short," or "right," whichever happened to be the case. Thus, this group always knew the direction of its error. In fact, all of the additional factors present for the Told-with-Punishment Group were present here, but the bell was absent. Or again, it might be said that the factors of knowledge concerning the performance were present, but that there was nothing to bring in the nature of punishment or reward, that is no qualitative implications were present in the knowledge, just a matter-of-fact statement of long, short, or right.

While the Knowledge Group receives all of the information concerning its performance that is received by the Told-with-Punishment Group, but does not have the bell, it receives more knowledge concerning its performance than do the Punishment and the Reward Groups. The Punishment and the Reward Groups receive only the bell, and so do not receive any information concerning the direction of their errors.

The striking fact is that while the results for the Knowledge Group show improvement, its improvement is nearly reliably less than that made by the Punishment and the Reward Groups, and is distinctly reliably less than that made by the Guess-with-Punishment and the Told-with-Punishment Groups. Thus, the bell, that is incentive in the nature of reward or punishment, is far more effective than even more accurate knowledge concerning the performance.

This is in agreement with the results of Johanson,¹⁰ who found that mere knowledge as to performance, while effective, was not as effective as an incentive stimulus which very definitely consisted in something to be avoided—electric shock. This is what we would call incentive in the nature of punishment.

It would seem that the "second" response which the subject makes to the physical stimuli, in reacting to them as something to be avoided or to be sought after, namely the punishment or the reward, facilitates the performance of the task at hand far more than simple facts about the performance. It follows that the individual must be responding to more cues, or perhaps more adequate cues, in the situation, than he does without the added factor of punishment or reward. But it would seem from this study, that the special cues given more attention, or made more effective, in the case of discriminating lengths, is not the knowledge of the direction of the error.

4. THE VARIATION IN PERFORMANCE AMONG THE GROUPS DURING THE FIRST FIVE TRIALS OF THE SECOND DAY

The average performance of each subject during the fifty trials of the first day, during which no incentive stimuli were introduced, was used as a base against which to measure the change in performance effected by the introduction of the incentive stimuli on the second day. The performance of each subject during any series of five trials was always considered as a percentage of this base. This seems to be the most suitable measure to use as a base, since it is derived from the greatest number of trials made without an incentive stimulus and so is probably the best measure of the subject's normal performance.

However, the first five trials of the second day were given before the incentive stimuli were introduced, to permit the subject to regain his general orientation toward the task at hand. If the average performance during the fifty trials of the first day is a good index of the subject's normal performance, it might be expected that the performance on this first series of the second day would not vary much from it. Examination of Table 2, which presents the group index of each series of five trials as a percentage of the average of the fifty trials of the first day, reveals that while four of the groups did not vary particularly from their average performance of the first day, the Reward and Guess-with-Punishment Groups made average errors of only 79% and 75%, respectively, of their average errors of their fifty trials of the first day. Two questions naturally arise. First: Is the variation

shown by these two groups greater than the chance variation which might be expected from day to day? Second: Is not the effect of the incentive stimuli employed in these two groups being overstated, since during the second day the average errors made by the R Group were actually reduced from only 79% to 26% of the average errors of the first day, and in the case of the GwP Group from only 75% to 15%, instead of from approximately 100% as in the case of all the other groups?

In order to throw some light on these questions and to safeguard any conclusions which might be drawn from the experiment, the data have been recalculated using the average error of the first five trials of the second day as the base. In Table 36, found on page 38 in the appendix, is shown the average error of the last five trials of each subject as a percentage of the average error of his first five trials of the second day. A summary of these by groups is found in Table 3A, below.

TABLE 3A
AVERAGE ERROR OF THE LAST FIVE TRIALS OF THE INCENTIVE SERIES AS A PERCENTAGE OF THE AVERAGE OF THE FIRST FIVE TRIALS OF THE INCENTIVE SERIES.

<i>Group</i>	<i>C</i>	<i>P</i>	<i>R</i>	<i>GwP</i>	<i>TwP</i>	<i>K</i>
Percentage	155	24	36	25	25	60

The reliability of the differences between the groups is shown in Table 4A, below.

TABLE 4A
RELIABILITY OF THE DIFFERENCES BETWEEN THE EFFECTS OF THE DIFFERENT INCENTIVES, AS MEASURED BY THE AVERAGE ERROR OF THE LAST FIVE TRIALS AS A PERCENTAGE OF THE AVERAGE OF THE FIRST FIVE TRIALS OF THE INCENTIVE SERIES

<i>Groups</i>	<i>Diff.</i>	<i>Sigma Diff.</i>	<i>Diff.</i> <i>Sigma Diff.</i>
C and P	131	24.0	5.5
C and R	119	24.6	4.8
C and GwP	130	23.8	5.5
C and TwP	130	24.4	5.3
C and K	95	27.5	3.5
P and R	12	8.5	1.4
P and GwP	1	5.6	0.2
P and TwP	1	7.8	0.1
P and K	36	15.0	2.4
R and GwP	11	7.7	1.4
R and TwP	11	9.5	1.2
R and K	24	15.9	1.5
GwP and TwP	0	7.1	0.0
GwP and K	35	14.6	2.4
TwP and K	35	15.6	2.2

Comparison of Tables 3 and 3A, and of Tables 4 and 4A, reveals that the use of the first five trials of the second day as a base against which to measure change effected by the incentives, instead of the average of the fifty trials of the first day, does not affect the relationships between the performances of the various groups. The only noteworthy change brought about by the use of the new point of reference is that the differences between the K Group and each of the other incentive groups are not so reliable. But this change is not sufficiently marked to affect any of the interpretations or conclusions.

V. EFFECT OF INCENTIVES ON MAKING ONE LINE EQUAL TO ANOTHER

In view of the fact that earlier work on the visual discrimination of length, which has been performed in the interest of the determination of difference thresholds, or in the study of psychophysical methods, has for the most part consisted in the setting of one length equal to another, we present here the results of an earlier experiment performed by the writer,⁷ in which incentive stimuli similar to those described above were employed when the task required of the subject was to set one length *equal* to an exposed standard.

The experimental set-up was like that described in Section II above, except that, in place of an illuminated glass bar in a dark room, a standard Galton Bar was used under ordinary conditions of illumination, and no time records were taken. The procedure and the incentive stimuli employed were the same, except that forty-five instead of fifty trials were made on each of the two days of the experiment, and that the subject's task was to set the length on the right of the dividing line *equal* to the 200 mm. length which was exposed as a standard on the left. The distance of the eyes from the bar was 1830 mm.

Three groups of five subjects each were used, consisting in a Control Group, a Punishment Group, and a Reward Group. The results of this experiment are summarized in Table 8, in a form comparable to Tables 2 and 3.

TABLE 8.

AVERAGE ERROR OF THE LAST FIVE TRIALS OF THE INCENTIVE SERIES AS A PERCENTAGE OF THE AVERAGE OF THE FORTY-FIVE PRACTICE TRIALS.

<i>Group</i>	<i>Subjects</i>				
Control	L.C. 86	J.E. 178	T.F. 91	O.K. 95	A.J. 189
Punishment	J.N. 18	J.J. 48	E.H. 115	A.R. 22	S.B. 24
Reward	A.H. 30	R.H. 11	M.M. 6	L.G. 33	M.T. 26

During the extent of the experiment, three subjects of the Control Group, as compared with those of the other two groups made a very small decrease in their average errors,

and in the case of two subjects, J.E. and A.J., the average errors showed a marked increase. But when the bell rang after each failure, as in the case of the Punishment Group, marked improvement resulted, except in the case of E.H., upon whom the incentive stimulus seemed to have no effect. The average errors of the other four subjects were brought down to from 18 to 48% of the average errors made by them during their practice trials. When the bell rang after each success, as in the case of the Reward Group, marked improvement resulted on the part of all the subjects. The average errors of their last five trials ranged from 6 to 33% of the average errors of their forty-five practice trials.

Thus we find these two conditions of incentive bringing about marked improvement in the visual discrimination of length, as compared with the change in performance during the same number of trials without incentive stimuli, when the task was that of making one length *equal* to another.

In the light of the previous work of this sort, it is of interest to compare the accuracy attained by these three groups, by considering the value of the average error in relation to the length of the standard. This comparison is made in Table 9.

TABLE 9
AVERAGE ERROR OF THE LAST FIVE TRIALS OF THE INCENTIVE SERIES AS A PERCENTAGE OF THE STANDARD LENGTH.

<i>Group</i>	<i>Subjects</i>				
Control	L.C. 4.6	J.E. 8.2	T.F. 4.1	O.K. 2.1	A.J. 7.2
Punishment	J.N. 0.9	J.J. 1.9	E.H. 8.2	A.R. 0.8	S.B. 1.0
Reward	A.H. 1.1	R.H. 0.5	M.M. 0.35	L.G. 0.8	M.T. 1.4

In the same general way that the results of the previous investigations of the visual discrimination of horizontal lengths by the method of average error have been compared with one another in Table 1, the data of Table 9 may be compared with that of Table 1. That is, one must bear in mind the variations in apparatus and other conditions of the experiments, which render impossible any real comparison. We note that in the case of the individuals reported in Table 1 the relationship of the average error to the standard length,

200 mm., ranges from 0.65 to 3.38%, while the performance of the five individuals in the Control Group (without incentive) in Table 9 ranges from 2.1 to 8.2%. In the case of the individuals who received incentive the range is from 0.35 to 1.9%, exclusive of E.H. who was not affected by the incentive stimulus. The greater error made by the subjects working without incentive than made by the earlier investigators is doubtless due to the differences in the conditions of the experiments. A very striking one is the difference in the distance from the eyes to the lengths being judged, which is from $2\frac{1}{4}$ to 6 times as great as those used in the other studies. In view of the relationships just pointed out, the relative errors made by the subjects who received incentive may be considered very small indeed, two of them being smaller in absolute difference from any in Table 1, while all are smaller than the relative error of five of the individuals whose work is reported in the latter table. Another difference in conditions of experimentation is that all of the earlier investigators made half of their trials with the comparison distance to the right and half to the left of the standard, while the individuals of Table 9 made all of their trials with the comparison distance to the right of the standard. However, the summary tables of Volkmann and of Appel (see 4) show that while the position of the standard affects the constant error, it has no effect upon the average error.

The subjects in the writer's control groups, who received no incentive stimuli, made no improvement with practice. This is true both of the task of making the comparison length twice as long as the standard, as shown in Table 12, and of the task of making the comparison length equal to the standard, as is shown in Table 10. It would be interesting if we could compare these data with those of other investigators. Only one, however, presents any data which bear upon this point, the others giving only their final summary tables. Kiesow¹² gives us the average error of each twenty of the hundred trials made by his subject. Table 11 shows these figures when the standard length was 200 mm. He also obtained no improvement with practice.

TABLE 10.

CONTROL GROUP.
 WHEN TASK WAS TO ADJUST ONE LENGTH EQUAL TO A 200 MM. STANDARD.
 THE RESULTS OF TWO DAYS PRACTICE WITHOUT INCENTIVE.
 AVERAGE ERROR.....IN MM.

<i>Subject</i>		<i>L.C.</i>	<i>J.E.</i>	<i>T.F.</i>	<i>O.K.</i>	<i>A.J.</i>
Series of five trials						
	1st	19.4	5.4	15.6	4.0	11.6
P	2nd	12.6	9.0	9.8	2.0	7.8
R	3rd	9.4	10.4	8.6	4.4	7.4
A	4th	9.0	9.6	8.8	4.0	5.2
C	5th	10.6	7.2	9.0	3.8	6.6
T	6th	5.0	5.2	9.6	5.6	3.8
I	7th	8.8	5.8	6.8	5.4	6.6
C	8th	12.2	15.6	4.4	4.0	10.2
E	9th	9.6	15.0	8.4	6.4	9.2
	1st	14.8	5.0	8.8	6.0	10.8
C	2nd	7.4	5.8	6.0	4.2	11.2
O	3rd	8.2	6.8	8.4	4.4	11.8
N	4th	6.4	15.0	8.6	4.2	10.0
T	5th	9.2	19.4	8.4	3.8	10.8
R	6th	9.2	17.8	8.6	6.4	10.8
O	7th	8.2	21.0	8.2	4.8	18.2
L	8th	7.2	17.8	6.8	3.2	16.0
	9th	9.2	16.4	8.2	4.2	14.4

TABLE 11.

TABLE FROM KIESOW SHOWING THE AVERAGE ERROR OF EACH TWENTY
 TRIALS WHEN THE STANDARD LENGTH WAS 200 MM.

1st 20 trials	1.275
2nd 20 trials	1.335
3rd 20 trials	1.33
4th 20 trials	1.26
5th 20 trials	1.33

VI. SUMMARY AND CONCLUSIONS

1. Mere continued repetition of this simple discrimination brought practically no change in accuracy.

2. Incentive stimuli effected improvement in accuracy in the discrimination of length, in the task of making one line equal to a standard, and of making it twice the length of a standard.

3. Under the conditions of incentive as described, visual discrimination of horizontal lengths by the method of average error was brought to a point of greater accuracy than had heretofore been attained.

4. The attitudes of reward and punishment created by our incentive stimuli were not reliably different in their effectiveness.

5. The physical stimuli used in creating the incentives do not in themselves constitute the incentives, for one and the same stimulus may at different times act as a quite different incentive. The incentive itself, then, may be an interpretation placed upon the stimulus by the subject, or a part of the subject's reaction to the stimulus, and so would depend upon the conditions under which the physical stimulus is given.

6. The experiment would indicate that in effecting improvement in discrimination of length the constant error is decreased first, even though no indication of the direction of the error is given; and is later followed, if the improvement is carried far enough, by decrease in variability, or variable error.

7. The time consumed in making a discrimination did not bear a consistent relationship to the accuracy of the discrimination.

8. Upon being informed that an error had been made, subjects were fairly well able to guess its direction; but this ability did not seem to play an important part in bringing about improvement.

9. Definite knowledge as to direction of error when given along with the punishment attitude did not effect more improvement than the reward or punishment attitude alone.

10. Definite knowledge when given in the absence of the reward or punishment attitude did not effect as much improvement as either of these attitudes when given without such knowledge.

11. The emotional nature of the reward or of the punishment attitude seemed to serve to heighten the attention to the task at hand, and perhaps enabled the subject to employ cues not otherwise brought into play.

VII. BIBLIOGRAPHY.

1. Arps, G. F.: Work with Knowledge of Results vs. Work Without Knowledge of Results. Psychol. Monog. 1920, 28, No. 3.
2. Chodin, A.: Ist das Weber-Fechnersche Gesetz auf das Augenmass Anwendbar? Arch. f. Ophthalmol. 1876, 23, 92ff.
3. Dodson, J. D.: Relative Values of Reward and Punishment in Habit Formation. Psychobiology 1917, 1, 231-276.
4. Fechner, G. T.: Elemente der Psychophysik. Leipzig 1889.
5. Fischer, R.: Grossenschatzungen im Gesichtsfeld. Arch. f. Ophthalmol. 1891, 37, 97ff.
6. Gates, G. S. and Rissland, L. O.: The Effect of Encouragement and Discouragement Upon Performance. Jour. of Educ. Psychol. 1923, 14, 21-26.
7. Hamilton, H. C.: The Effect of Incentives on the Accuracy of Judgment. Unpublished Master's Essay in the Department of Psychology, Columbia University, 1926.
8. Hügler, H.: Experimentelle Prüfung der Psychophysischen Methoden im Bereiche des Raumsinnes der Netzhaut. Philosophische Studien. 1892, 7, 232-297.
9. Hoge, M. A. and Stocking, R. J.: A Note on the Relative Value of Punishment and Reward as Motives. Jour. Animal Behav. 1912, 2, 43-50.
10. Johanson, A. M.: Influence of Incentive and Punishment on Reaction Time Arch. Psychol. 1922, 8, No. 54.
11. Judd, C. H.: Practice Without Knowledge of Results. Psychol. Rev. Monog. Supplements. 1905, 7, 185-189.
12. Kiesow, F.: Über die Vergleichung linearer Strecken und ihre Beziehung zum Weberschen Gesetze. Arch. f. d. ges. Psychol. 1926, 56, 421-451.
13. Münsterberg, H.: Beiträge zur experimentellen Psychologie. Freiburg 1889, 2, 150ff.
14. Rexroad, C. N.: Administering Electric Shock for Inaccuracy in Multiple Choice Reactions. Jour. Exper. Psychol. 1926, 9, 1-19.
15. Spencer, L. T.: The Effects of Practice Without Knowledge of Results. Amer. Jour. of Psychol. 1923, 34, 107-111.
16. Stephanowitsch, J.: Untersuchung der Herstellung der subjectiven Gleichheit bei der Methode der mittleren Fehler unter Anwendung der Registriermethode. Psychol. Studien. 1913, 8, 77-116.
17. Thorndike, E. L.: The Law of Effect. Amer. Jour. Psychol. 1927, Washburn Commemorative Volume, 212-222.
18. Warden, C. J. and Aylesworth, M.: The Relative Value of Reward and Punishment in the Formation of a Visual Discrimination Habit in the White Rat. Jour. Comp. Psychol. 1927, 7, 117-127.
19. Wright, W. R.: Some Effects of Incentives on Work and Fatigue. Psychol. Rev. 1906, 13, 23-24.

VIII. APPENDIX

TABLE 18.
CONTROL GROUP
CONSTANT ERROR..IN MM.

[illegible]

TABLE 19.
PUNISHMENT GROUP
CONSTANT ERROR.....IN MM.

<i>Subject</i>	<i>I.T.</i>	<i>F.K.</i>	<i>M.W.S.</i>	<i>D.H.</i>	<i>F.Gr</i>	<i>E.M.H.</i>	<i>H.R.K.</i>	<i>D.G.</i>	<i>V.L.</i>	<i>M.M.</i>
Series of five trials										
P R A C T I C E	2nd, 3rd, 4th	— 8.9	— 4.5	+ 85.5	+ 18.9	+ .6	— 11.7	+ 9.9	— 8.3	— 1.5 + 15.9
	5th, 6th, 7th	+ 3.0	+ 15.1	+ 81.1	+ 8.5	+ 4.4	— 8.8	+ 7.3	+ .1	— 8.5 + 25.2
	8th, 9th, 10th	— 10.5	+ 12.0	+ 87.7	+ 23.6	+ 10.0	— 8.1	+ 6.1	— 8.9	+ 6.5 + 22.0
P U N I S H M E N T	2nd, 3rd, 4th	+ 6.3	+ 10.5	+ 27.1	+ 8.2	— 4.1	— .1	+ 15.9	— 4.4	— 1.8 + 8.4
	5th, 6th, 7th	+ 10.1	+ 9.9	+ 6.1	+ .3	0.0	— 3.2	+ 9.2	— 3.0	— 2.1 — .3
	8th, 9th, 10th	+ 1.5	— .4	+ .1	+ 6.9	— 3.7	— 1.9	— 4.9	— .5	+ .4 + 3.5

TABLE 22.
TOLD-WITH-PUNISHMENT GROUP
CONSTANT ERROR..... IN MM.

Subject		<i>M.I.</i>	<i>E.G.</i>	<i>E.Re</i>	<i>J.Po</i>	<i>E.We</i>	<i>I.M.</i>	<i>E.S.</i>	<i>E.Ro</i>	<i>M.C.S.</i>	<i>B.C.</i>
Series of five trials											
P R A C T I C E	2nd, 3rd, 4th	-14.3	+1.9	-1.9	+15.9	+ 3.3	-16.6	-3.3	-14.5	- .8	+ 7.5
	5th, 6th, 7th	- 5.4	-3.7	+ .7	+11.1	+20.7	- 4.8	-3.9	-14.9	-6.0	+16.8
	8th, 9th, 10th	+ 3.9	- .5	-1.7	+12.1	+10.4	+ 4.3	-3.9	-15.2	-9.0	+ 7.5
P U N I S H M E N T	2nd, 3rd, 4th	- 8.2	- .4	- .6	+ 1.8	+ 4.1	- 3.6	+7.9	- 5.7	- .3	+ 1.1
	5th, 6th, 7th	- 4.6	+4.5	- 3	+ 3.1	+ .5	+ .1	-2.5	- 4.2	-1.7	+ 1.5
	8th, 9th, 10th	- 1.5	+1.2	+ .9	- .1	- .9	+ .5	-3.1	- 2.2	- .7	- .7

TABLE 24
CONTROL GROUP

<i>Subject</i>		<i>H.Q.</i>	<i>S.O.</i>	<i>I.L.</i>	<i>V.K.</i>	<i>I.R.</i>	<i>B.E.</i>	<i>H.A.K.</i>	<i>M.R.</i>	<i>M.M.S.</i>	<i>C.G.</i>
Series of five trials											
P R A C T I C E	2nd, 3rd, 4th	6.2	6.4	8.4	11.5	11.2	5.7	11.0	8.6	7.1	5.8
	5th, 6th, 7th	5.2	6.3	8.7	15.4	10.9	3.4	8.1	9.2	6.2	6.9
	8th, 9th, 10th	5.1	8.2	11.2	9.6	9.7	7.3	12.7	9.0	7.8	8.6
C O N T R O L	2nd, 3rd, 4th	7.0	9.3	9.7	4.6	6.2	4.0	11.7	9.0	4.7	6.8
	5th, 6th, 7th	7.3	9.3	11.1	12.9	4.6	4.9	9.2	3.2	4.6	7.0
	8th, 9th, 10th	5.1	10.7	12.9	13.4	12.8	5.8	4.6	6.8	5.1	6.5

TABLE 27
GUESS-WITH-PUNISHMENT GROUP
VARIABLE ERROR..... IN MM.

Subjects		<i>E.Be</i>	<i>I.B.</i>	<i>L.R.</i>	<i>B.M.</i>	<i>S.S.</i>	<i>L.L.</i>	<i>H.B.</i>	<i>F.Ga</i>	<i>V.C.</i>	<i>A.G.</i>
Series of five trials P R A C T I C E	2nd, 3rd, 4th	10.2	13.9	5.2	6.8	4.9	2.8	7.7	14.4	10.0	8.1
	5th, 6th, 7th	8.3	7.0	4.7	5.8	7.8	4.7	8.3	9.4	3.1	8.9
	8th, 9th, 10th	6.3	14.0	6.9	7.0	8.9	3.7	8.3	4.5	5.8	6.1
P U N I S H M E N T G U E S S w i t h	2nd, 3rd, 4th	6.2	11.9	4.7	9.1	7.4	2.8	6.0	15.9	5.6	4.3
	5th, 6th, 7th	7.9	11.4	4.8	3.2	8.0	3.0	12.9	7.9	8.0	5.5
	8th, 9th, 10th	4.6	6.6	5.5	7.2	2.0	1.7	3.7	7.2	4.6	4.3

TABLE 28
TOLD-WITH-PUNISHMENT GROUP
VARIABLE ERROR.....IN MM.

<i>Subject</i>	<i>M.I.</i>	<i>E.G.</i>	<i>E.Re</i>	<i>J.Po</i>	<i>E.We</i>	<i>I.M.</i>	<i>E.S.</i>	<i>E.Ro</i>	<i>M.C.S.</i>	<i>B.C.</i>
Series of five trials										
P R A C T I C E										
2nd, 3rd, 4th	9.6	4.8	4.3	11.8	6.2	6.7	10.5	8.2	5.9	11.7
5th, 6th, 7th	9.1	6.2	6.4	10.2	7.5	8.8	11.4	7.9	7.3	11.0
8th, 9th, 10th	7.2	5.6	6.2	4.8	7.3	11.2	7.7	8.3	5.3	12.1
P U N I S H M E N T										
2nd, 3rd, 4th	7.0	4.7	6.8	7.0	6.9	12.6	5.5	6.4	8.8	6.4
5th, 6th, 7th	4.7	5.1	2.8	4.0	5.7	7.7	7.2	7.9	6.2	6.8
8th, 9th, 10th ~	5.6	3.4 [*]	2.1	1.6	3.0	3.2	3.8	4.0	3.5	4.3

TABLE 31
PUNISHMENT GROUP
AVERAGE TIME PER TRIAL IN SECONDS FOR EACH SERIES OF FIVE TRIALS

<i>Subject</i>	<i>I.T.</i>	<i>F.K.</i>	<i>M.W.S.</i>	<i>D.H.</i>	<i>F.Gr</i>	<i>E.M.H.</i>	<i>H.R.K.</i>	<i>D.G.</i>	<i>V.L.</i>	<i>M.M.</i>	<i>Average</i>
Series of five trials											
P	29.3	59.0	10.6	27.2	29.6	17.5	12.6	11.7	15.9	19.5	23.3
2nd	29.5	43.4	9.0	25.9	23.9	13.6	15.0	9.0	12.9	15.3	19.8
3rd	27.1	43.7	...	27.2	27.7	11.9	17.3	10.9	13.6	14.8	21.6
R	27.1	30.0	...	24.2	23.0	13.7	16.9	12.8	11.6	14.7	18.3
4th	24.7	37.0	6.9	28.1	21.6	9.6	17.9	11.2	12.6	14.0	18.4
C	27.0	31.1	8.3	22.0	21.8	8.9	16.0	12.7	12.8	12.3	17.3
6th	24.7	33.5	8.2	26.3	21.1	11.4	17.5	11.6	13.7	10.7	18.9
T	22.3	33.0	8.4	21.2	23.0	11.4	18.2	11.8	12.3	12.1	17.4
I	24.8	36.4	8.4	26.9	16.8	11.9	19.2	11.0	12.0	11.8	17.9
C	24.3	32.2	8.4	...	23.2	10.1	16.1	14.0	10.9	9.8	16.6
10th											
P	25.0	29.3	8.3	21.5	13.2	12.6	18.9	13.4	11.1	13.6	16.7
2nd	27.6	...	8.4	33.0	14.5	14.9	16.4	13.7	11.3	17.1	17.4
3rd	26.2	29.5	7.4	20.7	17.2	13.2	20.6	12.4	9.4	17.7	17.4
4th	27.7	25.9	7.0	22.5	12.9	15.6	24.6	12.7	9.1	14.6	17.3
5th	24.2	27.6	7.2	23.4	16.7	16.9	23.4	12.8	9.5	14.6	17.6
H	24.4	28.4	6.7	20.6	18.3	16.7	14.5	12.7	10.5	14.1	16.7
6th	24.9	21.7	7.4	15.9	23.1	18.2	14.0	15.1	9.6	14.2	16.4
M	22.0	17.6	7.0	15.9	19.3	15.1	13.3	15.0	11.7	13.6	15.1
E	21.3	18.1	7.0	17.0	19.9	16.8	12.4	14.9	9.0	15.1	15.2
N	18.0	15.2	8.7	16.9	17.6	17.9	11.0	15.6	8.6	13.3	14.3
10th											
Average Fifty Practice Trials	26.1	37.9	8.6	25.4	23.2	12.0	16.7	11.7	12.8	13.5	18.8

TABLE 32
REWARD GROUP
AVERAGE TIME PER TRIAL IN SECONDS FOR EACH SERIES OF FIVE TRIALS

Subject	C.L.	H.P.	J.Pi	H.U.	E.Wa	H.R.	C.T.	V.F.	H.S.	S.R.	Average
Series of five trials											
1st	41.9	33.5	22.2	57.7	44.8	13.6	17.7	17.1	51.0	28.4	32.8
2nd	40.9	16.5	23.0	45.0	66.5	10.2	12.4	12.8	37.3	16.3	28.1
3rd	25.2	20.2	32.5	61.4	45.3	12.7	14.1	11.3	30.5	19.8	27.3
4th	26.7	17.6	27.6	38.4	61.3	11.6	14.7	12.6	28.0	...	26.5
5th	34.1	24.5	30.1	33.5	44.5	13.3	12.8	16.3	27.1	20.1	25.6
6th	28.2	18.0	24.5	25.0	37.5	12.5	12.5	17.8	18.8	16.1	21.1
7th	24.2	21.1	24.6	35.0	40.1	11.9	14.1	17.1	16.9	14.9	22.0
8th	24.4	23.5	22.5	31.6	50.0	10.9	13.9	16.1	13.8	14.9	22.2
9th	24.6	20.2	23.5	31.8	41.1	9.0	14.2	22.3	15.3	21.0	22.3
10th	24.9	12.7	19.6	35.4	47.7	10.1	14.0	15.1	16.6	15.9	21.2
1st	27.1	8.6	19.7	48.0	38.6	11.2	13.3	18.3	24.9	17.5	22.7
2nd	27.8	11.3	27.4	25.6	37.3	11.0	16.7	18.5	15.4	10.5	20.2
3rd	22.6	11.4	25.5	24.6	50.8	10.4	16.5	13.9	13.4	7.9	19.7
4th	21.0	13.4	21.7	20.1	41.0	10.7	21.2	18.3	10.6	8.8	18.7
5th	23.9	12.1	18.1	21.9	44.8	10.2	16.6	12.2	9.3	9.2	17.8
6th	27.4	12.4	19.2	22.3	26.2	12.5	16.3	14.6	10.9	9.5	17.1
7th	25.7	10.2	19.2	26.6	34.7	11.4	16.1	12.3	9.9	12.3	17.8
8th	21.3	12.1	19.6	22.7	30.9	12.9	15.2	23.0	10.8	12.6	18.1
9th	23.7	13.5	19.4	25.0	38.0	14.8	17.7	23.6	10.3	14.2	20.5
10th	29.7	14.2	17.7	13.9	16.4	21.5	10.7	11.3	16.9
Average Fifty Practice Trials	29.5	20.8	25.0	39.5	47.9	11.6	14.0	15.9	25.5	16.7	24.6

TABLE 33
GUESS-WITH-PUNISHMENT GROUP
AVERAGE TIME PER TRIAL IN SECONDS FOR EACH SERIES OF FIVE TRIALS

Subject		<i>E.Be</i>	<i>I.B.</i>	<i>L.R.</i>	<i>B.M.</i>	<i>S.S.</i>	<i>L.L.</i>	<i>H.B.</i>	<i>F.Ga</i>	<i>V.C.</i>	<i>A.G.</i>	<i>Average</i>
Series of five trials P R A C T I C E	1st	23.6	37.5	24.6	25.0	14.7	59.9	12.3	27.9	19.1	26.4	27.1
	2nd	12.5	32.5	20.1	..	10.3	61.3	27.9	31.6	10.2	15.6	24.7
	3rd	20.4	27.5	27.7	16.0	8.9	55.4	18.7	38.8	11.7	16.4	24.2
	4th	15.4	25.9	24.6	17.0	8.9	57.6	22.6	34.1	10.0	13.5	23.0
	5th	14.1	26.2	25.4	13.3	8.8	44.6	18.9	35.3	10.6	11.5	20.9
	6th	11.5	22.2	21.4	..	8.6	30.0	14.5	32.9	9.6	12.5	18.1
	7th	12.0	21.2	23.1	19.0	8.5	46.2	10.3	32.9	11.7	12.8	19.8
	8th	11.2	21.6	18.4	15.8	7.6	48.6	9.0	37.3	10.2	12.4	19.2
	9th	16.0	19.6	23.2	17.0	8.0	43.6	8.0	23.1	10.1	12.1	18.1
	10th	15.5	15.4	17.9	13.0	8.3	48.0	9.1	23.4	9.4	11.3	17.1
G U E S S w i t h P U N I S H M E N T	1st	16.8	31.7	25.8	16.5	11.6	...	10.6	23.1	13.0	10.8	17.8
	2nd	20.1	24.7	27.9	18.8	17.6	47.8	10.2	31.8	18.2	10.9	22.8
	3rd	15.4	28.5	29.3	23.0	15.8	51.4	9.0	22.3	17.8	11.6	22.4
	4th	9.3	25.9	24.7	20.6	21.0	66.9	10.0	17.4	20.7	13.4	23.0
	5th	9.0	23.7	24.6	21.6	20.4	60.5	9.5	12.1	20.8	13.6	21.6
	6th	16.4	22.2	20.9	22.4	21.2	54.9	9.8	10.2	12.4	14.5	20.5
	7th	9.8	20.9	19.7	28.6	17.8	58.3	10.0	9.7	17.1	15.8	20.8
	8th	10.7	17.9	23.6	26.7	19.3	42.4	9.8	9.4	16.7	15.9	19.2
	9th	8.9	21.0	19.4	30.7	20.7	35.0	8.9	9.6	21.1	16.4	19.2
	10th	8.6	16.6	...	29.4	21.1	33.1	8.0	10.9	21.9	16.3	18.4
Average Fifty Practice Trials		15.2	25.0	22.6	17.0	9.3	49.5	15.1	31.7	11.3	14.5	21.1

TABLE 35
KNOWLEDGE GROUP
AVERAGE TIME PER TRIAL IN SECONDS FOR EACH SERIES OF FIVE TRIALS

<i>Subject</i>	<i>M.K.</i>	<i>E.V.</i>	<i>L.W.</i>	<i>E.H.H.</i>	<i>P.H.</i>	<i>E.M.</i>	<i>J.F.</i>	<i>E.B.j</i>	<i>M.J.</i>	<i>T.L.</i>	<i>Average</i>
Series of five trials	1st	11.2	31.4	27.0	20.3	12.6	24.6	10.8	6.9	13.3	10.4
P	2nd	9.7	21.4	29.3	28.6	9.4	31.2	13.1	6.4	10.8	11.7
R	3rd	9.4	21.1	27.5	20.1	10.0	26.5	15.1	6.9	10.0	11.3
A	4th	9.5	14.4	27.1	19.6	9.2	20.4	14.7	6.7	13.4	8.9
C	5th	8.3	15.5	25.2	16.7	9.7	22.2	16.2	7.8	12.2	10.1
T	6th	8.4	13.8	29.7	15.8	8.5	21.8	13.7	7.6	8.4	7.6
I	7th	8.9	13.2	28.8	14.9	9.8	25.9	13.9	7.4	9.5	8.2
C	8th	9.2	12.3	25.1	14.1	10.9	30.8	12.9	6.9	9.2	8.5
E	9th	8.6	11.8	33.4	13.9	9.3	25.0	12.3	6.9	9.4	8.2
	10th	7.9	10.3	31.2	17.7	11.0	20.5	11.9	7.8	10.1	9.1
K	1st	8.6	17.2	22.3	16.7	10.6	15.8	15.5	5.4	7.8	8.8
N	2nd	9.4	18.0	26.9	13.2	9.8	26.8	15.6	5.8	7.7	14.2
O	3rd	10.0	16.6	23.2	12.3	8.5	22.1	11.9	7.0	8.8	14.5
W	4th	8.8	16.1	24.8	11.1	8.8	26.2	11.7	6.8	8.9	13.5
L	5th	9.2	14.6	26.1	10.3	9.6	30.5	9.8	6.2	7.6	14.0
E	6th	10.9	12.6	19.9	15.1	9.0	36.4	10.0	6.1	8.0	13.8
D	7th	8.5	12.4	24.5	15.9	10.0	31.7	10.1	5.3	8.8	14.2
G	8th	8.5	11.0	22.9	15.0	9.7	33.2	10.8	6.4	9.5	14.3
E	9th	8.8	10.7	22.5	10.9	8.9	33.0	9.7	5.9	9.0	18.3
	10th	8.8	12.4	19.6	13.0	8.9	20.9	12.0	6.1	9.7	17.2
Average Fifty Practice Trials		9.1	16.5	28.4	18.2	10.0	24.9	13.5	7.1	10.6	9.4

TABLE 36
AVERAGE ERROR OF THE LAST FIVE TRIALS AS A PERCENTAGE OF THE AVERAGE OF THE FIRST FIVE TRIALS OF THE INCENTIVE SERIES

Group	Subjects										Average
	H.Q.	S.O.	I.L.	V.K.	I.R.	B.E.	H.A.K.	M.R.	M.M.S.	C.G.	
Control	217	111	271	89	205	100	33	98	184	242	155
Punishment	I.T. 23	F.K. 19	M.W.S. 4	D.H. 18	F.Gr 36	E.M.H. 14	H.R.K. 56	D.G. 21	V.L. 38	M.M. 9	24
Reward	C.L. 56	H.P. 35	J.Pi 16	H.U. 18	E.Wa 90	H.R. 40	C.T. 41	V.F. 38	H.S. 11	S.R. 18	36
Guess-with-Punishment	E.Be 13	I.B. 35	L.R. 12	B.M. 18	S.S. 38	L.L. 15	H.B. 31	F.Ga 26	V.C. 39	A.G. 23	25
Told-with-Punishment	M.I. 16	E.G. 18	E.Re 9	J.Po 14	E.We 15	I.M. 80	E.S. 25	E.Ro 9	M.C.S. 35	B.C. 31	25
Knowledge	M.K. 48	E.V. 18	L.W. 39	E.H.H. 11	P.H. 38	E.M. 45	J.F. 171	E.Bj 46	M.J. 78	T.L. 104	60

VITA

Born, Cedar Rapids, Iowa, March 6, 1903. Graduated from Washington High School, Cedar Rapids, Iowa, 1921. B.A., Cornell College, Mount Vernon, Iowa, 1925. Graduate work, Columbia University, 1925-28. M.A. 1926. Lydia Roberts Fellow, 1925-27. Assistant in Psychology, Columbia University, 1927-28. Sigma Xi, 1928. Associate Member of the American Psychological Association. Instructor in Psychology, Temple University, 1928—.

AN INVESTIGATION INTO THE VALIDITY OF NORMS WITH SPECIAL REFERENCE TO URBAN AND RURAL GROUPS

BY
MYRA E. SHIMBERG

Submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy, in the Faculty of Philosophy,
Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, EDITOR

No. 104

NEW YORK ·
March, 1929

To My Parents
SIMON JAMES SHIMBERG
and
ESTHER LIGHT SHIMBERG
With Gratitude and Affection

ACKNOWLEDGMENTS

Acknowledgments are gratefully made to Dr. W. Healy and Dr. A. Bronner of the Judge Baker Foundation, Boston, with whose help this experiment was started; to Dr. F. Dunn for valuable criticism and unfailing cooperation; to Miss H. H. Heyl and Mr. A. H. Cochrane for their assistance in securing rural and urban subjects; to Mrs. E. Tunnell for aid in the preparation of the bibliography; to Professor H. E. Garrett for helpful suggestions, especially with reference to statistical procedure; and to Professor R. S. Woodworth, under whose guidance this investigation was completed.

The writer wishes to thank the Superintendents whose names are mentioned in the text and the teachers who are too numerous to mention individually, for their enthusiastic support of this project.

Acknowledgment is also appreciatively made to Wellesley College which, through the Alice Freeman Palmer Fellowship, helped the writer to carry on this work.

CONTENTS

<i>Chapter</i>	<i>Page</i>
I. Introduction	5
II. Construction and Standardization of Information Tests A and B	7
III. Age and Sex Differences	19
IV. Differences Between Various Rural Schools	31
V. Rural and Urban Differences on Information Tests A and B	42
VI. The Nationality of the Subjects and the Corre- sponding Scores	63
VII. Summary and Conclusions	75
Appendix	77
Bibliography	82

An Investigation into the Validity of Norms With Special Reference to Urban and Rural Groups

CHAPTER I

INTRODUCTION

Mental testing requires no justification; it is now an indispensable part of the technique of the educational psychologist, the vocational expert and the psychopathologist. Whereas twenty-five years ago the available psychological tests could have been described in a pamphlet, nowadays whole books on tests fill our shelves and the supply seems inexhaustible.

This rapid growth is splendid in many respects, but it has led inevitably to hasty and inaccurate technique in many quarters. In a recent manual on tests (8), of which the present writer is a co-author, 126 rather widely used tests of a certain type were described. Of these at least one-fifth did not attain the low numerical level of sample which we had set as a minimum in standardization. Admitting that mere addition of numbers does not necessarily increase validity, it is, nevertheless, obvious that averages based on a sampling of the population that is less than fifty cases have little significance.

The applicability of norms, however, depends on many other factors besides the number of cases on which the tests are standardized. One has to consider the composition of the group from the standpoints of variability, sex, racial and national components, etc. Moreover, the original scaling of the test with the possible favoring of one group to the exclusion of others must be taken into account.

We present below the results of an analysis of our 126 tests in regard to the selection of subjects. Our analysis is based on the data included in the original articles. As stated above, a great many tests, 24.6%, were not well enough standard-

ized to pass even our low numerical requirements. The tests analyzed below represent, then, the upper 75% of our tests.

We believe that in the field of comparative psychology—in so far as it deals with the comparison of one human group with another—the result of the misuse of tests has been most marked. The asserted inferiority of any group over another may, perhaps, be laid to the tools utilized in the analysis rather than to any intrinsic differences in the groups themselves. No particular comment is necessary. These figures speak for themselves.

TABLE I

Sex	<i>Taken into Consideration</i> 60 8%	<i>Not Taken into Consideration</i> 39 2%	
σ , A D, P E.	<i>Given</i> 39 2%	<i>Not Given</i> 37 9%	<i>Not Given but Could be Figured From Given Data</i> 23 5%
	<i>Taken into Consideration</i>	<i>Not Taken into Consideration</i>	<i>Some Consider- ation Given</i>
Racial Composition	9 8%	88 2%	2 0%
Social Composition	11 8%	50 9%	37 3%
Numbers	<i>Large</i> 59 0%	<i>Fairly Large</i> 17 5%	<i>Adequate Only At Certain Ages</i> 23 5%

We have attacked this problem in the following manner: Having prepared two tests scaled and standardized according to the best known methods, we have used them in the comparison of two groups—urban and rural children. Our aim is twofold: 1. to examine the importance of the differentiation of norms according to sex, educational groups, localities, racial composition; 2. to inquire whether differences between our groups are a function of the group mentality or our tests as tools.

Each chapter, dealing with one of these topics, contains a summary of the pertinent literature. We have made no attempt to be exhaustive, since we have found it necessary to draw so many fields into our discussion.

Moreover, our research itself has touched only the fringes of a somewhat unexplored field. Our purpose will have been accomplished if it contributes to the more careful scrutiny of psychological articles on race and group differences.

CHAPTER II

THE CONSTRUCTION AND STANDARDIZATION OF INFORMATION TESTS A AND B

An information test of more than local import, adequately scaled and standardized, and adapted to clinical use, seemed to offer a virgin field for experimentation. Of the few information tests already in existence, none seemed to fill these requirements.

F. Kuhlmann and T. G. Foran were kind enough to send us the information tests which they use. Neither of these has, to our knowledge, been published and the standardization, if it exists, is unavailable. Doubtless similar material, in more or less embryonic form, is utilized in clinics all over the country.

For the most part, vocabulary tests similar to that in the Stanford-Binet have, despite certain shortcomings, been used as information tests. The three mentioned below have been so designated.

Whipple's "Range of Information" test is merely an extension of his vocabulary test, the words being so selected "that each shall be representative of some specific field of knowledge or activity."¹ From a performance one can ascertain the fields with which the subject is acquainted, *i.e.*, American history, golf, photography, etc. It is no adequate gauge of his commonsense information as a whole.

Pressey and Shively, in 1919 (2), utilized a test composed of 10 groups of 10 words each, representing 10 different fields. This test was designed to fit the needs of delinquents and since, moreover, it was loosely organized and not standardized, it need not concern us here.

Weeks, 1928 (5), constructed the Berkshire test, consisting of three sets of 50 words each (one taken from the Stanford-Binet), to be defined according to the multiple choice method, *e.g.*, pancreatic: nerve sweetbread universal panic-stricken.

We know of only two tests at all similar to our own:

¹ 6, p. 683.

Terman, in his "Genetic Studies of Genius" (3), mentions an individual information test devised for the comparison of average and gifted children. It is in part identical with the information section of the Stanford Achievement test, but is more reliable and covers a wider range. Each of the two forms contains 335 items, *e.g.*:

The earth is shaped most like a baseball, football, pear.

The house-fly spreads Bubonic plague, typhoid, yellow fever.

This is an extremely good test. But since it takes an hour to administer, it would have been impractical for our purposes even if it had been reported before we started our investigation. To our knowledge, no one besides Terman has published results on these tests, although they may be available in printed form. The test is not reported in full in "Genetic Studies of Genius."

Eastman, in 1926 (1), compiled an information test of 100 questions for use in the Wayne psychopathic clinic. While admirably adapted for the use for which it was intended, its questions are extremely local in scope, (*e.g.*, Name five car lines in Detroit), and thus not suited to our purpose.

The following paragraphs describe the construction and standardization of our tests A and B. We should like to emphasize the fact that A and B do not refer to alternative forms. Whenever the A test is referred to, it should be thought of as the test scaled on urban children, whereas the B test was scaled on rural children.

SCALING INFORMATION A

Construction. The first step in the construction of the test was the preparation of the material. In this preliminary work, we were fortunate enough to have the cooperation of a group of teachers of elementary and high schools. Questions which they submitted were recast, modified and supplemented until we had eighty questions² of varying degrees of difficulty, covering the different fields of commonsense information, *e.g.*, nature, industry, etc. They were so framed that they could not be answered by "yes" or "no." These questions were then mimeographed on two sheets (easy and hard questions being scattered throughout in irregular order) and tried out on 764 urban children in grades 4-12, inclusive. Not

² For test in detail, see Appendix.

more than two grades were taken from the same school. The distribution of cases follows:

TABLE II

<i>Grades</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>	<i>IX</i>	<i>X</i>	<i>XI</i>	<i>XII</i>
<i>No. of Cases</i>	71	96	102	101	98	111	32	59	94

Since we desired, in this preliminary tryout, to test all the questions and not the individual children, it was extremely important that each question be attempted by each child. We secured the cooperation of the teachers in having the test given over a period of two or three days, so that each section should be given due emphasis.

From the returns we calculated the % of correct responses for each question. These %'s were converted into S.D. units,³ as may be seen from Table III. Unsatisfactory questions (because of ambiguity in interpretation or difficulty in scoring) were discarded.⁴

From the remaining 72 questions, 25 questions were selected in such a manner that there was approximately the same distance in S.D. units between each 2 questions. It was thus possible to state that the third question was as much more difficult than the second, as the 25th question was more difficult than

³ 54 pp. 396-400. Table is based on area of probability curve, assuming base line to be broken off at $\pm 3.0\sigma$.

⁴ Some of the answers encountered in scoring the papers were sufficient reward for the labor involved. Some of these gems are shared with the reader.

11. How is it that newspapers can be sold for less than the cost of printing them? "Because people put in ads about divorces."
25. Name three precious stones. "Diamonds, coal, gold." "Diamond, rubi, Plymouth Rock."
26. Why is the moon light at night? "So the sun can rest." "Because it is made of cheese."
54. Where does Congress meet? "In the Gulf of Mexico."
55. What is a civil war? "Where 2 enemies fight with arms agreed on beforehand."
58. Why did the Pilgrims come to this country? "To get gold and precious stones."
59. How many sides has a triangle? "Any number of unequal sides."
64. What is the freezing point of water? "North and South Poles."
69. What is vaccination for? "To rest and go to see the world." "It is a permit mark on a child's arm." "To show you're old enough to go to school."
71. Name four general reasons that prevent a would-be immigrant from entering the United States. "Whisky, Pistols, Germs, Piosnin."
74. Name five insects. "Americans, Italians, Finnish, Swedish and Spaniard." "vilin, catar, piano, drum, float."

10 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

TABLE III
PRELIMINARY SCALE A

Questions Arranged in Order of Difficulty for Urban Children		
σ	% Passing	
72	99 0	What are the colors in the American flag?
82	98 7	Of what are shoes made?
1 06	97 5	How many cents are there in a quarter?
1 08	97 4	How many hours are there in a day?
1 19	96 6	What may we expect when we see heavy black clouds?
1 32	95 5	What holiday comes in December?
1 36	95.1	How do you know a policeman when you see him?
1 37	95 0	What is our national song?
1 45	94 1	Who is president of the United States?
1 47	93 8	To what public building can you go for books?
1 51	93 3	How much does it cost to mail a letter to any city in the U S ?
1 54	92 9	What people were in America when the white men came?
1 71	90 3	Of what is butter made?
1 74	89 8	What state do you live in?
1 77	89 2	What holiday do we now celebrate that was first celebrated by the Pilgrims?
1 88	87 1	How many sides has a triangle?
1 93	85 9	How old must you be before you can vote?
1 93	85 9	Of what is paper made?
1 94	85 8	How many months are there in a year?
1 94	85 8	Where does the sun rise?
1 95	85 5	How many states are there in the U.S.?
1 95	85 5	Name four different trees
1 97	85 1	Who was the first president of the U S ?
2 07	82 7	How many pints are there in a quart?
2 08	82 3	Name five vegetables
2 09	82 1	What do the stars in the American flag represent?
2 13	81 0	What is the capital of the U S ?
2 15	80 5	In your city, what is the youngest age at which a child can leave school?
2 18	79 7	What is the shortest month in the year?
2 19	79 3	What is the largest river in the U S ?
2 35	74 5	Why did the Pilgrims come to this country?
2 36	74 2	What is a submarine boat?
2 37	73 7	Who is the Governor of your State?
2 41	72 3	What is the largest city in the U S ?
2 47	70 5	For how many years is the president of the U S elected?
2 50	69 4	Who was the president of the U. S. during the World War?
2 55	67 7	Name three precious stones
2 57	66 9	Why should we kill flies?
2 58	66 7	Why is it dark at night?
2 67	63 2	Name a country in Europe which is a republic.
2 67	63 2	How many weeks are there in a year?
2 68	62 7	What three things do most plants need in order to live?
2 78	59 0	Why don't we see the stars in the day time?
2 79	58 6	About how often do we have a full moon?
2 79	58 6	Where does Congress meet?
2 81	57 7	How does the beating of your heart keep you alive?
2 83	57 1	What is the value of the smallest silver coin we use?
2 88	55 1	What causes an eclipse of the sun?
2 90	54 3	Why do we celebrate the 4th of July?
2 91	53 7	What artificial waterway connects the Atlantic with the Pacific?
2 94	52 5	Name the greatest English writer of plays?
2 96	51 8	What is steam?
2 98	51 1	What form of government have we in the U S.?

SPECIAL REFERENCE TO URBAN AND RURAL GROUPS 11

σ % Passing

3 05	48 0	How can banks afford to pay interest on the money you deposit?
3 06	47 7	What is the economic value of Alaska to the U. S.?
3 24	40 5	Name five insects.
3 31	37 7	Of what is rubber made?
3 33	37 0	What is a civil war?
3 38	35 0	Why is the moon light at night?
3 39	34 6	Name two stones used for building purposes.
3 41	34 0	Name five cities in the U. S. that have a population of over half a million.
3 47	31 9	Name the Great Lakes.
3 58	28 0	How is it that newspapers can be sold for much less than the cost of printing?
3 72	23 4	What is the freezing point of water?
3 88	18 8	What is the usual economic result of the over-production of any commodity?
4 02	15 4	In what country is Vienna?
4 12	13 0	In what month of the year do the days begin to grow shorter?
4 28	10 0	What are the functions of the three branches of our Government? (In three words or phrases)
4 39	8 1	Of what use are insects to flowers?
4 43	7 5	Name four general reasons that would prevent a would-be immigrant from entering the U. S.
4 82	3 3	What is the function of respiration?
5 19	1 3	What is a referendum in government?

the 24th. The test, as it appeared in its final printed form, is reproduced on the following page.⁵

Standardization. As subjects in the standardization of this test we secured the entire school population of a city of 47,876 inhabitants (1920 Census). This city was considered by its assistant superintendent to be "average." In a state survey in arithmetic it had fallen midway in the distribution.

The method of T. scaling, described in detail by McCall,⁶ was used in the construction and standardization of Scaled Information A. This probably gives more accurate results than either the percentile or age scales. McCall states that "the T scale is believed to be superior to any of the previously described methods. . . . It scales the total score. It employs the simple total. It allows each test element done to affect the scale score, thereby increasing reliability. Its units are equal in the generally accepted sense at all points on the scale."⁷

The tests were all administered on the same day under conditions as nearly standard as possible. Each teacher gave the test to her own pupils following precise instructions. Judging from our contact with superintendent and staff we are confident in stating that the conditions under which these

⁵ This test can be obtained from Stoelting & Co., Chicago, Ill.

⁶ For details, see 53, pp. 272-306.

⁷ 52, p. 96.

12 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

A SCALED INFORMATION TEST

1. What are the colors in the American flag?
2. Of what are shoes made?
3. How many hours are there in a day?
4. What holiday comes in December?
5. Who is President of the United States?
6. What people were in America when the white men came?
7. What State do you live in?
8. Where does the sun rise?
9. Name five vegetables.
10. What is the largest river in the United States?
11. Why did the Pilgrims come to this country?
12. For how many years is the President of the United States elected?
13. What three things do most plants need in order to live?
14. How does the beating of your heart keep you alive?
15. How can banks afford to pay interest on the money you deposit?
16. Name five insects.
17. What is a civil war?
18. Name five cities in the United States with a population of over half a million.
19. How is it that newspapers can be sold for less than the cost of printing?
20. What is the freezing point of water?
21. What is the usual economic result of the over-production of any commodity?
22. What are the functions of the three branches of our government? (In three words or phrases).
23. Name four general reasons that will prevent a would-be immigrant from entering the United States.
24. What is the function of respiration?
25. What is a referendum in government?

NAME _____
SCHOOL _____
AGE Yrs _____ Mos _____
GRADE _____

group tests were given were as ideal as possible. In all, 6477 usable records were secured.

The scoring was done by two trained workers.⁸ After correcting a large number of these papers a key was formulated,⁹ which contained acceptable answers, common non-acceptable answers, etc. All the papers were then graded according to these principles. Since the test had been so well scaled, it was practicable to give one point for each correct answer. No partial credits were allowed. Each pupil's score was the number of questions answered correctly.

⁸ One was the author. The other was Miss Marjorie Meehan, to whom the writer is extremely indebted for her painstaking and reliable work.

⁹ See Appendix.

The exact procedure as outlined by McCall for T scaling was then followed. Papers from our 886 12-year-olds were separated from the others, and the percentage of 12-year-olds passing each question was determined. (See Table IV). The percentage of 12-year-olds exceeding plus half those reaching 0, 1, 2, . . . 25 questions was then computed. From these percentages the scale score (in S.D. values) was determined from the table furnished by McCall.¹⁰ Since the 12-year ability did not include scores as low as necessary, the scale was extended at the lower end by repeating the above processes for question 0 in the case of 10-year-olds. Since the scale score for question 0 was 3 points below their score for question 1, 3 points was subtracted from the 18 shown in Table IV. The same procedure for questions 22-25 was followed with 16-year-olds.

TABLE IV
T SCALING INFORMATION A

<i>Total No. Questions Correct</i>	<i>No. of 12-yr.-olds Pupils*</i>	<i>% Exceeding + Half Those Reaching</i>	<i>Scale Score</i>
0	0		15
1	1	99 94	18
2	1	99 83	21
3	6	99 43	25
4	2	98 98	27
5	8	98 42	29
6	15	97 12	31
7	29	94 64	34
8	37	90 91	37
9	49	86 06	39
10	69	79 40	42
11	92	70 32	45
12	96	59 71	48
13	108	48 19	50
14	113	35 72	54
15	83	24 65	57
16	62	16 47	60
17	45	10 44	63
18	30	6 21	65
19	19	3 44	68
20	12	1 69	71
21	9	51	76
22	0		79
23	0		82
24	0		85
25	0		91

*At upper level, above first line, scale score is based on performance of 10-yr.-olds. At lower level, below second line, it is based on performance of 16-yr.-olds.

¹⁰ 53, p. 279.

The test was then standardized for age and grade. A discussion of these results is contained in the ensuing chapters.

In order to see if the test was local in significance, *e.g.*, in some way favored our particular group of children, it was tried out on a group (106) of tenth grade children in Springfield, Illinois, and 71 tenth graders in Portland, Oregon. In neither case were the new groups at a loss. In fact, each group did somewhat better than the original. This seems to rule out the possibility of the test's being very localized in import.

It was necessary, also, to measure the reliability of the test. Since it was found impossible to give the test twice and no duplicate form existed we correlated the separate scores on odd and even questions, and calculated the reliability of the whole test by the formula $rx = \frac{2rh^{11}}{1 + rh}$. The self correlation of 360 cases chosen at random from all grades was .83, P.E.01. This comes within the range of minimum reliability coefficients, as set forth by Garrett.¹²

Scaled Information A is now in daily clinical use at the Judge Baker Foundation, Boston, Mass., and has been found satisfactory as a test of practical commonsense information.

SCALING INFORMATION A

In the course of our investigation, it became necessary to make a test for rural children as Information A had been constructed for urban children. As far as we know, there are no precedents for this. It has always been taken for granted that tests scaled and standardized on white city children should be made the basis of comparison with other groups, but seldom indeed have the tables been turned. Velma Helmer attempted to devise tests "to demonstrate the possibility of improving the Indians' relative score at verbal tests by presenting situations which are probably more familiar to the Indian"¹³ than those found in standard tests. Reference was made to "hogun," "tepee," Indian customs, materials, etc. This obviously favored the Indian children. No such attempt has, to our knowledge, previously been made with rural children.

Construction. A preliminary test of 80 questions,¹⁴ was

¹¹ 51, p. 271.

¹² *Ibid.*, p. 269.

¹³ 58, p. 42.

¹⁴ See Appendix.

compiled from material submitted by rural teachers. Thirty-seven questions were identical with questions in the preliminary A test. As in the case of the former test, the author framed all the questions and included none which appeared to be merely local in significance, or highly specialized. These questions were given to all the children in an entire rural district. This covered 52 schools and included 416 children from the 4th-12th grades. It must be remembered that rural schools consist, in great part, of one-room schools, averaging 10-25 pupils from all grades. The distribution of cases follows:

TABLE V

<i>Grades</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>	<i>IX</i>	<i>X</i>	<i>XI</i>	<i>XII</i>
<i>No of Cases</i>	49	57	68	73	32	60	33	29	15

The tests were given by the individual teachers as before. The exact instructions followed may be found in the appendix.

From the returns, the percentage of correct responses for each question was determined (see Table VI) in the way already described in the previous section and the final test of 25 questions secured in the same manner as before. It is reproduced below.

TABLE VI

PRELIMINARY SCALE B

Questions Arranged in Order of Difficulty for Rural Children		
σ	$\%$ <i>Passing</i>	
29	99 8	What may we expect when we see heavy black clouds?
84	98 6	How many cents are there in a quarter?
91	98 3	Of what is butter made?
1 37	95 0	How much does it cost to mail a letter to any city in the U. S.?
1 43	94 3	How can you keep milk from souring?
1 43	94 3	At what time of year do many leaves turn red?
1 45	94 1	What are the four seasons?
1 47	93 9	From what does maple sugar come?
1 57	93 5	What is our national song?
1 68	90 8	How many states are there in the U S ?
1 70	90 5	Name a vegetable that grows above ground
1 73	89 9	How many pecks are there in a bushel?
1 74	89 9	What kind of dairy cow gives the richest milk?
1 77	89 1	How many pints are there in a quart?
1 82	88 2	Name the young of the sheep, cow, horse.
1 85	87 7	Name five fruits
1 93	85 9	What do we mix with ice to help us freeze ice-cream more quickly?
1 93	85 9	What is the largest city in the U S ?

16 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

σ	%	Passing	
2 00	84 2		Who was the first President of the U. S.?
2 01	84 1		What is the capital of the U. S.?
2 01	84 1		Name four different trees
2.02	83 7		Draw a square and an oblong.
2 03	83 6		How old must you be before you can vote?
2 03	83 6		What do the stars in the American flag represent?
2 08	82 3		Why should we kill flies?
2 08	82 3		From what animal do we get mutton?
2 09	82 1		Of what is paper made?
2 11	81.4		Name two birds that stay North in the winter.
2.15	80 3		Of what is rubber made?
2 16	80 2		Why does seasoned wood burn more easily than green wood?
2.17	79 9		Why are crops hoed?
2 23	78 2		How many sides has a triangle?
2 25	77.4		Name five crops.
2 29	76.4		What holiday do we celebrate that was first celebrated by the Pilgrims?
2 30	75 9		Who is the Governor of your State?
2 38	73 3		What is the shortest month in the year?
2 42	72 1		What tree doesn't shed its leaves in the Fall?
2 46	70 7		Why is it necessary to limit the hunting season?
2 47	70 3		Tell one way of finding out the age of a tree
2 60	65 6		Why don't we see the stars in the daytime?
2 61	65 4		Why is it dark at night?
2 61	65 4		What kind of cloth is made from flax?
2 62	65 1		Name three states in the U. S. where cotton is raised.
2 64	64 1		Name a famous American inventor and tell what he invented
2 65	63 7		How many weeks are there in a year?
2 71	61 4		What is a submarine boat?
2 73	60 6		About how often do we have a full moon?
2 75	60 0		What artificial waterway connects the Atlantic with the Pacific?
2 80	58 2		Name a country in Europe which is a Republic.
2 82	57.2		What is the correct temperature for a living room?
2 84	56 4		Name five wild flowers
2 88	53 1		Where does Congress meet?
2 89	54 5		What form of government have we in the U. S.?
2 92	53 2		What is the economic value of Alaska to the U. S.?
2 94	52 7		What is steam?
2 95	52 2		Name the continents in order of size
3.01	49 3		Why does frost form on the inside of the window pane?
3 06	47 3		Name three products made from wheat
3 08	47 0		Name two animals that hibernate in winter.
3 08	46 5		Name three uses of forests.
3 11	45 3		What is the highest court in the U. S. called?
3 11	45 3		What causes an eclipse of the sun?
3.11	45 3		Name two stones used for building purposes.
3 17	43 3		Who was the President of the U. S. during the World War?
3 28	38 8		Give one reason for the rotation of crops.
3.32	37 2		How do sponges grow?
3 39	34 6		In what part of the day are the shadows longest?
3 42	33 7		Name three different plants from which sugar is made.
3.72	23 6		In what month of the year do the days begin to grow shorter.
3.84	20 0		Why is the moon light at night?
3.87	19 2		Name two differences between the barks of birch and oak trees.
4.20	11 4		How can you locate the Pole star?

Standardization. This test was then administered to 4875 rural children from eight districts sampling Northern, Southern, Eastern, Western and Central sections of the state. The

SPECIAL REFERENCE TO URBAN AND RURAL GROUPS 17

map outlines the districts included and shows how adequately the state was covered.

The tests were administered as before under precise instructions.¹⁵ The correcting was done by two trained workers,¹⁶ a key formulated,¹⁷ and the tests scored as before. The self correlation was found to be .84, P.E.01 (580 cases) for grades 4-8 inclusive. Norms for age, grade, sexes, union and one-room schools were secured. These will later be discussed in detail.

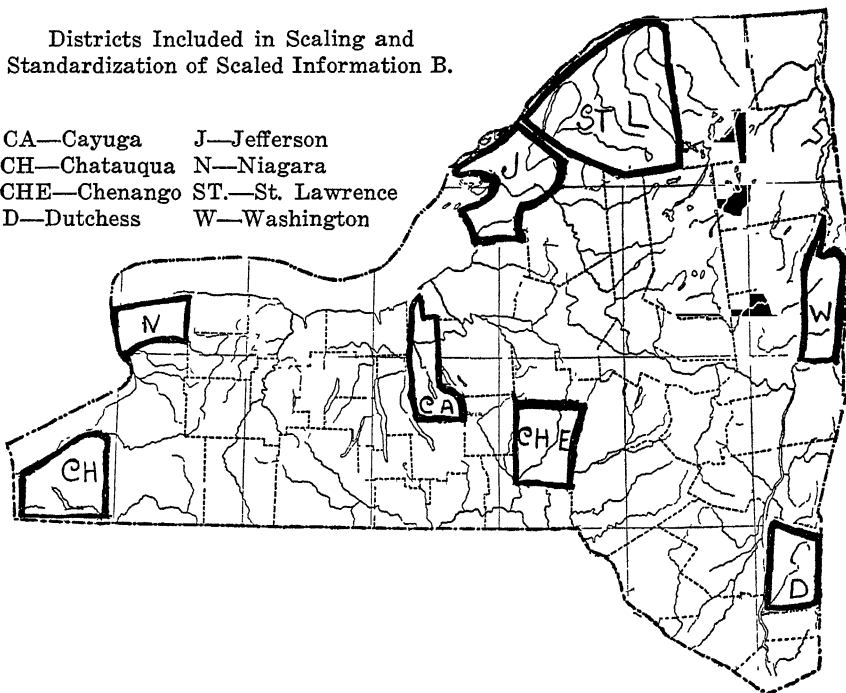
SUMMARY

1. Information Test A was constructed for clinical use. It was tried out on 764 urban children. It was T scaled and standardized on the entire school population (6477) of one city. The test was found to be statistically reliable.

2. Information Test B was scaled on 415 rural children

Districts Included in Scaling and
Standardization of Scaled Information B.

CA—Cayuga	J—Jefferson
CH—Chataqua	N—Niagara
CHE—Chenango	ST.—St. Lawrence
D—Dutchess	W—Washington



¹⁵ See Appendix.

¹⁶ Again the author, this time assisted by Miss A. Anastasi who is to be highly commended for the painstaking care with which she accomplished this work.

¹⁷ See Appendix.

18 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

from one district and standardized on 4875 rural children from eight districts in one state.

SCALED INFORMATION TEST—B

1. Of what is butter made? _____
2. How much does it cost to mail a letter to any city in the U. S.? _____
3. What are the four seasons? _____
4. What is our national song? _____
5. Name a vegetable that grows above ground. AGE Yrs _____ Mos. _____
6. Name the young of the sheep, cow, horse. GRADE _____
7. What do we mix with ice to help us freeze ice-cream more quickly? _____
8. Name four different trees.
9. From what animal do we get mutton?
10. Why does seasoned wood burn more easily than green wood?
11. Name five crops.
12. What tree doesn't shed its leaves in the fall?
13. Tell one way of finding out the age of a tree.
14. Why don't we see the stars in the daytime?
15. About how often do we have a full moon?
16. Name five wild flowers.
17. What is the economic value of Alaska to the U. S.?
18. Why does frost form on the *inside* of the window pane?
19. Name three uses of forests.
20. Who was the President of the U. S. during the World War?
21. Give one reason for the rotation of crops.
22. Name three different plants from which sugar is made.
23. In what month of the year do the days begin to grow shorter?
24. Why is the moon light at night?
25. How can you locate the Pole star?

CHAPTER III

AGE AND SEX DIFFERENCES

The previous chapter has dealt with the construction and process of standardization of Tests A and B. In this and the following chapters we shall consider in detail the results of trying these tests out on an urban and rural population.

Age and Grade Scores on Information A. Table VII shows the grade and age norms of the urban population on Test A. These norms are T scaled, so that 50 (in this case 50.1) is the average for 12-year-olds. It will be observed that the averages increase perceptibly from age to age and grade to grade.

TABLE VII
SCALED INFORMATION A
Urban Scores

<i>Age</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>	<i>Grade</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>
9	371	42.2	9 3	4	970	40 7	9 0
10	768	43.6	9 3	5	988	47 8	9 0
11	883	47.5	9 3	6	955	50 2	9 0
12	886	50 1	10 0	7	932	54 7	8.1
13	952	54 7	10 1	8	822	58 6	7 9
14	838	57 4	9 3	9	638	60 9	7 8
15	614	59.4	9 1	10	505	61 2	8 1
16	530	61.7	8 8	11	389	65 5	8 5
17	325	64 5	10 4	12	278	68 9	9 1
18	198	68 8	8 6	*College Freshmen	93	81 8	5 2
				*College Seniors	69	84 9	5 3

*These are not included in the age groups.

Table VIII shows us that the differences between consecutive ages and grades are entirely reliable. Conventional reliability is obtained when $\frac{\text{Difference}}{\text{Sigma "}} = 3$. The $\frac{\text{Differences}}{\text{Sigma "}}$ in these cases average 5.5 for age, and 8.0 for grade. We may say, then, that we have a test highly discriminative for age and grade levels. This, of course, makes it very useable in clinical practice.

It has been found that where adequate measures are used, the rate of mental growth can be shown to be fairly uniform.

TABLE VIII
 SCALED INFORMATION A
 Reliability of Differences—Urban Schools
 AGE

<i>Years</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff. is above 0</i>
9 & 10	1 4	.59	2 4	99.2 in 100
10 & 11	3 9	.47	8 3	100 in 100
11 & 12	2 6	.55	5 5	100 in 100
12 & 13	4 6	.48	9 6	100 in 100
13 & 14	2 7	.46	5.9	100 in 100
14 & 15	2 0	.49	4.1	100 in 100
15 & 16	2 3	.53	4.3	100 in 100
16 & 17	2.8	.69	4 1	100 in 100
17 & 18	4 3	.87	4 9	100 in 100

GRADE				
<i>Grades</i>				
4 & 5	7 1	.40	17.8	100 in 100
5 & 6	2 4	.40	6 0	100 in 100
6 & 7	4 5	.39	11 2	100 in 100
7 & 8	3.9	.39	10 0	100 in 100
8 & 9	2.3	.42	5 5	100 in 100
9 & 10	3	.48	6	74 in 100
10 & 11	4 3	.56	7 7	100 in 100
11 & 12	3 4	.69	4 9	100 in 100

Brooks (9), for example, has demonstrated by retests of children with a large battery of tests that growth is very nearly constant from 9-15. More nearly pertinent to our own findings is that of Terman (with the information test described in Chapter II), who reports that the age norms "gave an approximately straight line from ages 8-15."¹⁸ Weeks, on her vocabulary information test, also reports that "the average of grade scores showed a continuous increase."¹⁹

Besides the 6,265 urban children, a few college students were tested. Their averages are included in Table VII. While the test was obviously too easy for them, still it did discriminate between the Freshmen and Seniors.²⁰

¹⁸ 3, p. 299.

¹⁹ 5, p. 62.

²⁰ Dr. R. Brotemarkle (of the University of Pennsylvania) and the author have prepared an information test overlapping with Test A to the extent of five questions, the other 20 being scaled in a similar fashion for college students. Dr. Brotemarkle has used this scale in rather extensive tests of college students. His article, with results, is to be published shortly.

Graph I illustrates the regularity and consistency of the age and grade norms. Their deviation from a straight line drawn from start to finish would not be very marked. Also, the norms for age and grade are unusually consistent in regard to each other. This would undoubtedly be more often encountered in standardization work if large enough, unselected samples of the population were taken.

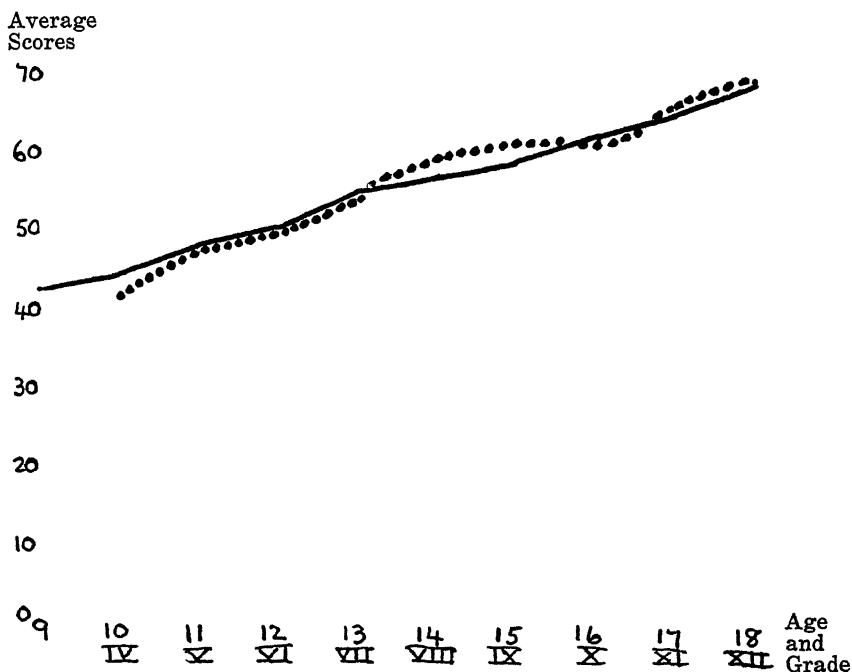


FIG. 1. Information A. Age and Grade Scores

Graph II shows the distribution of scores for sixth grade children. Its close approximation to the curve of normal probability is evident, and once more points to the value, or, in fact, sheer necessity, for accurate standardization, of using a large number of cases.

AGE AND GRADE SCORES ON INFORMATION B

Table XVII (in chapter IV) shows the age and grade averages on Scaled Information B for boys and girls, from one-room and union schools. The age averages for the one-room schools progress smoothly and consistently until age 15

22 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

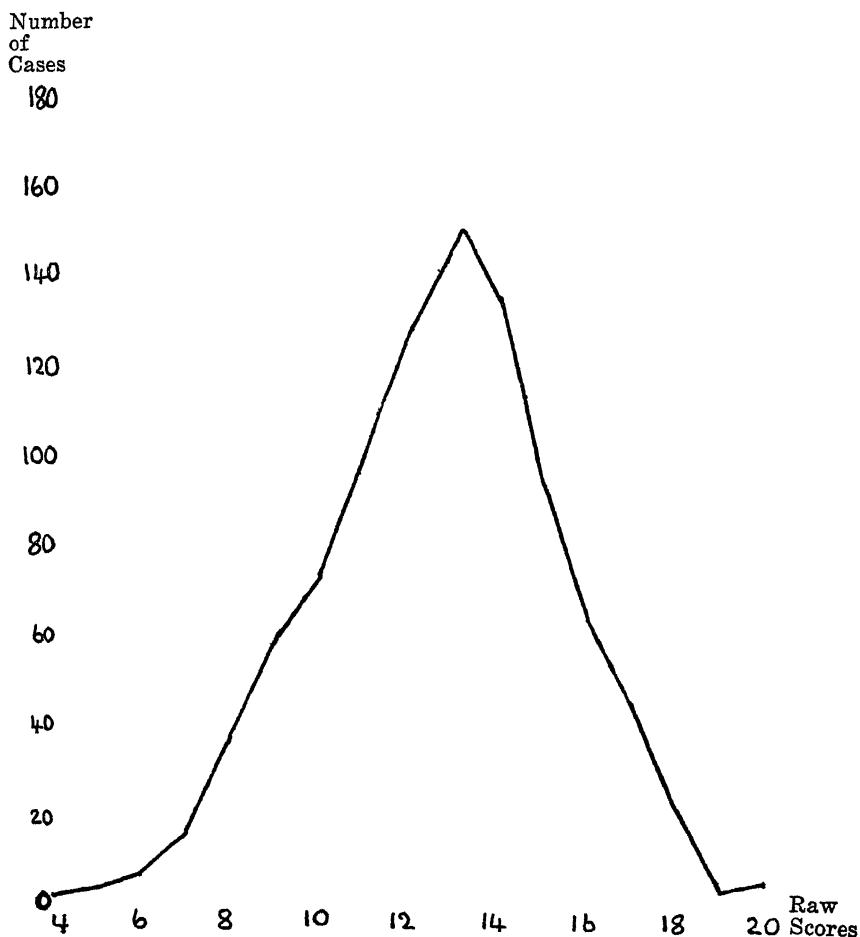


FIG. 2. Information A. Grade VI. Distribution

when there is a slump. This is readily explainable. The one-room schools proceed only as far as the 8th grade. At this juncture, pupils have to change to Union schools for the completion of their education. Consequently only the dullest children will be left in the one-room schools at the upper ages.

Graph III is plotted similarly to Graph I. While there is a reasonably consistent straight line relationship, the age and grade norms, however, are not as readily superimposable as in Test A. This is probably due to the fact that grade placings are much more flexible in rural schools. When 25 children from grades 2-8 are gathered together in one school room, it

is reasonable to conclude that their grade placings will be less rigid and more reliant on the criteria of the individual teacher than in the city schools.

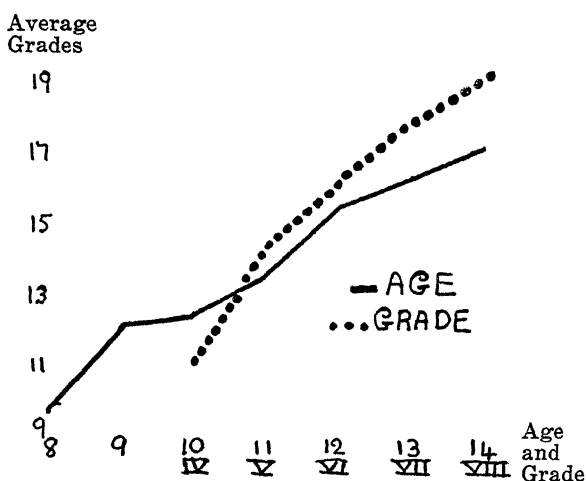


FIG. 3

Information B. Age and Grade Averages

For the union schools, there is a steady increase in the norms from ages 9-18 except for the age 13, which, for the girls, is the same as for year 12. This is not true of the boys' norms for those years, and we are at a loss to explain this one inconsistency. At no age, however, is there any slump.

As far as the grade norms go, progression from 4th to 12th is consistent and fairly smooth for the boys and also for the girls with one exception of a slight slump at grade 12.

Table IX shows the reliability of the differences from age to age and grade to grade. They are not as marked in this test as in Information A. However, the chances for a true difference average, in the case of age, 93.9 for one-room schools, and 87.3 for union schools; in the case of grade, 96.5 for one-room schools and 92.7 for union schools. These are rather significant differences.

Sex Differences. It is extremely regrettable that when the norms for Scaled Information A were compiled, no distinction was made between those for boys and girls. There was plenty of precedent for this lumping procedure. For example, in all the Pintner Paterson norms no attention is paid to

TABLE IXa
 SCALED INFORMATION B
 RELIABILITY OF AGE AND GRADE DIFFERENCES—
 ONE-ROOM SCHOOLS
 BOYS

<i>Years</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff. is above 0</i>
8 & 9	2 44	1 06	2 30	98 9 in 100
9 & 10	40	77	52	69 in 100
10 & 11	1 08	46	2 35	99 2 in 100
11 & 12	1 94	41	4 73	100 in 100
12 & 13	92	.39	2 36	99 2 in 100
13 & 14	82	36	2 28	98 9 in 100
14 & 15	.38	40	95	83 in 100
<i>Grades</i>				
3 & 4	2 60	68	3 82	100 in 100
4 & 5	3 22	32	1 00	84 in 100
5 & 6	2 08	30	6 93	100 in 100
6 & 7	1 44	30	4 80	100 in 100
7 & 8	1 40	30	4.67	100 in 100
GIRLS				
<i>Years</i>				
8 & 9	2 86	1 01	2 83	100 in 100
9 & 10	56	53	1 06	85 in 100
10 & 11	1 92	42	4 57	100 in 100
11 & 12	68	.44	1 55	93 5 in 100
12 & 13	1 04	40	2 60	99 5 in 100
13 & 14	94	41	2 29	98 9 in 100
14 & 15	62	47	1 32	90 in 100
<i>Grades</i>				
3 & 4	98	96	1 02	84 in 100
4 & 5	3 04	.35	8 68	100 in 100
5 & 6	2 06	37	5 57	100 in 100
6 & 7	2 00	35	5 71	100 in 100
7 & 8	60	32	1 88	97 in 100

sex differences. In such performance tests, where boys have so often been shown to excel, we should consider such a procedure exceedingly dubious. In comparing girls the averages are probably too high, for the boys, too low. In the 21 form-board and construction tests we analyzed, wherever there was sex differentiation (*i.e.*, in only 5 cases), a considerable superiority on the part of the boys was consistently found.

As far as verbal tests go, however, sex differences have even more rarely been taken into account. Even in the splendid standardization of the Thorndike McCall scale, so adequate as to scaling and number of cases, norms are not given for boys and girls separately.

TABLE IXb
 SCALED INFORMATION B
 RELIABILITY OF AGE AND GRADE DIFFERENCES—
 UNION SCHOOLS
 BOYS

<i>Years</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff. is above 0</i>
9 & 10	2.70	1.02	2.65	99.6 in 100
10 & 11	.82	.69	1.19	88 in 100
11 & 12	1.26	.60	2.10	98 in 100
12 & 13	2.36	.57	4.14	100 in 100
13 & 14	.44	.50	.88	82 in 100
14 & 15	.64	.42	1.43	92 in 100
15 & 16	.40	.43	.93	83 in 100
16 & 17	.64	.52	1.23	89 in 100
17 & 18	.90	.52	1.73	96 in 100
<i>Grades</i>				
4 & 5	3.20	.49	6.53	100 in 100
5 & 6	1.66	.42	3.95	100 in 100
6 & 7	1.30	.45	2.89	100 in 100
7 & 8	1.36	.40	3.40	100 in 100
8 & 9	.84	.33	2.55	99.4 in 100
9 & 10	.80	.37	2.16	98.3 in 100
10 & 11	.30	.82	.37	64.5 in 100
11 & 12	.48	.82	.59	73 in 100
<i>GIRLS</i>				
<i>Years</i>				
9 & 10	.94	.79	1.19	88 in 100
10 & 11	1.58	.59	2.68	99.6 in 100
11 & 12	1.74	.59	2.95	99.8 in 100
12 & 13	0			
13 & 14	1.36	.47	2.89	99.8 in 100
14 & 15	.90	.48	1.88	96 in 100
15 & 16	.46	.50	.92	82 in 100
16 & 17	1.12	.45	2.49	99.4 in 100
17 & 18	.36	.42	.86	80 in 100
<i>Grades</i>				
4 & 5	3.24	.42	7.71	100 in 100
5 & 6	2.68	.40	6.70	100 in 100
6 & 7	.12	.42	.29	62 in 100
7 & 8	1.34	.41	3.27	100 in 100
8 & 9	.90	.35	2.57	99.5 in 100
9 & 10	.66	.41	1.61	94 in 100
10 & 11	1.12	.42	2.67	99.6 in 100

Whipple, in a recent article, says: "The outcome of any intelligence test may be regarded as conditioned by a series of factors, such as age, sex, native ability, school training, practice and race. Of these factors comparatively little attention has been paid to sex, perhaps because the sex difference was regarded as of little magnitude or of little practical import. In the classification of pupils, for example, the scores of either

sex have been, *doubtless quite properly* (italics are ours), compared with the single standard score for the age or grade in question, which has been itself derived from the scores obtained by hundreds of children of both sexes combined, to furnish the standard age score or the standard grade score."²¹

Our results seem to cast grave doubt on that "doubtless quite properly." For, in test B,, the norms are carefully differentiated as to sex for both urban and rural children, and the results are remarkably clear cut. Graphs V-VII show that for grades 4-7²² the dotted line representing the girls' averages runs below and practically parallel to the continuous

TABLE X
COMPARISON BETWEEN BOYS AND GIRLS—Information B
Reliability of Differences—One-Room Schools

<i>Years</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff. is above 0</i>	
(Favor of Boys)					
8	1 10	1 20	92	82	in 100
9	.68	.84	81	79	in 100
10	.52	.42	1 24	89	in 100
11	-.32	.46	- 70	76	in 100 (Girls)
12	.94	.39	2 41	99 2	in 100
13	.82	.40	2 05	98	in 100
14	.70	.37	1 89	97	in 100
15	.46	.49	.94	83	in 100
16	.42	.68	62	73	in 100
<i>Grades</i>					
3	-.92	1.12	-.82	79	in 100 (Girls)
4	.72	.35	2 06	98	in 100
5	.88	.32	2 75	99.7	in 100
6	.90	.30	3 00	100	in 100
7	.34	.28	1 21	88	in 100
8	1 14	.33	3 45	100	in 100
Reliability of Differences—Union Schools					
<i>Years</i>					
9	- 68	1 10	- 62	73	in 100 (Girls)
10	1 18	.69	1 61	94	in 100
11	.42	.61	.69	76	in 100
12	- 06	.59	- 10	54	in 100 (Girls)
13	2 30	.51	4 50	100	in 100
14	1 38	.46	3 00	100	in 100
15	1 08	.45	2 40	99 2	in 100
16	1 02	.49	2 08	98	in 100
17	.54	.48	1 13	87	in 100
18	1 08	.47	2 30	98 9	in 100

²¹ 18, p. 111.

²² Only these grades were included in order to allow a comparison with the urban group, where only grades 4-7 were tested.

line, representing the boys' averages. At no place do these two lines cross. When we examine Tables X and XI, however, it appears, at first sight, that the difference is not entirely consistently in favor of the boys. For the one-room schools, the 11-year girls and the third grade girls exceed the boys. For the union schools, the 9- and 12-year girls exceed the boys. The average $\frac{\text{Difference}}{\text{Sigma "}}$ in these cases is .56 (71 chances in 100 for a true difference), which, of course, is very low. The other differences are in favor of the boys, the average number of chances in 100 for a true difference being 91.2 for one-room schools and 94.2 for union schools. For the urban group, the difference is consistently in favor of the boys and the average number of chances in 100 for a true difference is 94.5.

TABLE XI
COMPARISON BETWEEN BOYS AND GIRLS (2)—Information B
Reliability of Differences—Union Schools

<i>Grades</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff is above 0</i>
(Favor of Boys)				
4	.82	51	1 61	94 in 100
5	1 28	40	3 20	100 in 100
6	26	42	62	73 in 100
7	1 44	45	3 20	100 in 100
8	1 46	36	4 06	100 in 100
9	1 40	32	4 38	100 in 100
10	1 54	46	3 35	100 in 100
11	72	80	90	82 in 100
12	1.36	.48	2 83	99.7 in 100
Reliability of Differences—Urban Schools				
<i>Years</i>				
9	1.18	.98	1 20	88 in 100
10	.88	.49	1 80	96 in 100
11	1 10	.57	1 93	97 in 100
12	1 28	.53	2 42	99 2 in 100
13	1 34	.76	1 76	96 in 100
<i>Grades</i>				
4	.80	.40	2 00	98 in 100
5	98	.51	1 92	97 in 100
6	1 74	.47	3.70	100 in 100
7	.46	.58	.79	79 in 100

The 11- and 12-year divergence of the norms is not surprising. It has been found often that around or just before the period of adolescence, the balance is turned in favor of one sex for a time, although merely temporarily. The third grade and 9-year divergence may perhaps be explained on the

grounds of the paucity of cases. We had only 38 boys and 24 girls of the third grade from one-room schools and 20 boys and 33 girls from the union schools. This is the only place where we have any small number of cases, and, under no conditions, as we have often emphasized, can such numbers give positively reliable results. So, we may fairly conclude that, with the possibility of an exception for a short pre-adolescent period, our boys are consistently superior to our girls.

There is no reason to believe, moreover, that the content of the test particularly favored the boys. There were even a few more girls than boys in the original group on which the test was scaled. Table XII shows the comparison of 100 boys and 110 girls from District A on test B. By a strange coincidence, the girls exceed in 12 questions, the boys in 12 questions, there being an equal score on the first question. The average percentage by which the boys exceed the girls is 8.8%, while that of the girls exceeding the boys is 6.3%. These figures account for the age and grade superiority of the boys. It doesn't seem, however, that with the questions roughly stacking as we have indicated above there could have been any great selection in favor of the boys except in so far as a random group of information questions would always tend to favor them. Our results seem to indicate a genuine sex difference.

For a detailed analysis of the great mass of contributions to the study of sex differences, the reader is referred to the elaborate summaries so carefully compiled, at frequent intervals, by such authorities as Woolley (20), Hollingworth

TABLE XII
QUESTIONS FAVORING EACH SEX—Information B

Question	% in Boys' Favor	% in Girls' Favor	Question	% in Boys' Favor	% in Girls' Favor
1	0	0	14	18	
2		2	15		6
3		8	16		10
4		19	17	6	
5	8		18	6	
6	2		19		5
7	7		20	10	
8		3	21		3
9		3	22	8	
10	11		23		9
11	7		24	7	
12		6	25		1
13	15				

(12, 13), Goodenough (10), etc. Lincoln (15) has lately contributed an entire book on this subject.

Goodenough, comparing recent studies says, "The most outstanding impression which one gains . . . is the inconsistency of the various findings."²³ In the field of general information, however, there is some agreement, and it is gratifying to discover that our results are confirmed by those of several other investigators, whose work is summed up below:

The Berlin Child Study Association (17), studying 2,238 children just entering school, found that the number of concepts grasped by the boys exceeded that of the girls. Hall (11), in his study on Boston children, also found this to be true.

Pressey, testing 2,544 school children, finds that the "boys show a superior ability on the arithmetic test and a slightly higher average on the test for practical information."²⁴

King and McRory (14), testing Freshmen with Whipple's Range of Information, found that 61.6% of the boys reached the median of the girls.

Terman, in his study of gifted children, found that the "gifted boys excel gifted girls in general information, arithmetic and spelling."²⁵

Book and Meadows (7), testing 2,422 boys and 3,503 girls (all H. S. seniors) with the Pressey Mental Survey Scale, found that the boys were slightly superior to the girls on the whole test. The tests in which the boys most exceeded the girls were those in arithmetical ability and practical information. In commenting on the fact that Mrs. Pressey, testing subjects 9-15, found girls somewhat superior on the test as a whole, Book and Meadows say: "The superior rating of the girls is clearly due to the fact that the girls' development is accelerated from 1 to 2 years during this period."²⁶ It is interesting that the four cases in our own results where the girls excel the boys are 12 years or below, and in the 12-year group there is practically equality.

Goodenough, surveying the literature, says: "The weight of the evidence, therefore, seems to point to the conclusion that boys have acquired a truly wider range of general information than have girls by the time they arrive at the beginning of the grammar school period."²⁷

²³ 10, p. 441.

²⁴ 16, p. 333.

²⁵ 3, p. 306.

²⁶ 7, p. 77.

²⁷ 10, p. 453.

Eastman used an information test of 100 questions with juvenile delinquents. She finds the boys superior to the girls. "It may be presumed that boys, being less restricted in their freedom on city streets, acquire a more extensive knowledge of their immediate environment."²⁸

This falls in line with our own findings. It is true that most of our children didn't have city streets to roam, but it is true in the country as in the city that the boys have greater freedom than the girls. They are constantly out and doing, rubbing elbows with newcomers, and gathering bits of information from sundry sources, whereas the girls are kept indoors to a considerable extent, with domestic duties, etc. We are not surprised, then, to find that the boys' range of information is wider and more inclusive than that of the girls.²⁹

SUMMARY

1. Information A was found to be highly discriminative for age and grade. The sigmas of the difference averaged 5.5 for age and 8.0 for grade.

2. Information B was found to be highly discriminative for age and grade up to age 15. An explanation for the apparent discrepancy at the upper ages is offered in the text.

3. The distribution of scores was found to resemble closely that of the curve of normal probability.

4. The various graphs illustrate the consistency and reliability of our norms and distributions. This is believed to be due to the use of large numbers and adequate statistical procedure. This is urged as a requisite for all standardizations.

5. A clear cut sex difference favoring the boys was demonstrated.

6. Minute analysis of 210 papers of boys and girls led to the conclusion that the content of the test did not unduly favor the boys.

7. A review of the pertinent literature shows that, on the whole, our results are in line with those of other investigators.

8. We urge that all norms be sex differentiated when used for individual appraisement.

²⁸ *1*, p. 208.

²⁹ In view of the fact that the New York rural survey showed the boys to be considerably more retarded than the girls, this sex difference in information is particularly interesting. The rural schools are probably better adapted to the girl than to the boy. Most of the one-room schools, at least, are taught by women who tend naturally to stress female occupations.

CHAPTER IV.

DIFFERENCES BETWEEN VARIOUS RURAL SCHOOLS

The 4875 rural children who acted as subjects in our final B test were recruited from eight districts as shown on the map in Chapter III. In all, approximately 300 schools, one-room and union, were reached. This included 2330 children in one-room schools, 1808 from union schools, and 737 unclassified.

The one-room school "is that institution of the open country employing a single teacher and providing elementary education (*i.e.*, up to 8th grade only) for the children in the area in which it is located."³⁰ The average number of children in each school is approximately 25.

According to the government report on consolidated schools, the typical union school "is the result of uniting five districts or schools and abandoning four schoolhouses. . . . It serves an area of 36 square miles. . . . The school is organized on the 8-4 plan, enrolling 204 children in the elementary grades, 76 in the high school. . . . The teaching staff of 11 persons, including the superintendent, is divided on a basis of six or seven in the elementary grades and five or four in the high school. . . . It transports 110 (43%) of the children enrolled an average of 4.7 miles one way in 35 minutes."³¹

This chapter deals with two kinds of comparison, *i.e.*, 1. between the various localities tested; 2. between the two types of schools—one-room and union.

COMPARISON BETWEEN RURAL SCHOOLS IN DIFFERENT DISTRICTS

Tables XIII-XVI show the age and grade averages for the one-room and union schools in each separate district. The averages are, of course, not very reliable because they are necessarily based on such small numbers. For example, few one-room schools have more than two or three children at the upper grades, so that when one has canvassed all the schools in the district, the cases for each grade are still extremely limited.

³⁰ 26, p. 339.

³¹ 21, pp. 4-5.

Bearing this in mind, I think an inspection of the tables will reveal that there are no marked and consistent differences between the eight localities. Such variations as there are may doubtless be laid to the small numbers of cases. No district can be said to be consistently better than any other. The averages cross and re-cross each other and often end up the same. For example, taking 12-year one-room girls, there is a difference of only .58 between the highest and lowest average in the seven districts. This is confessedly one of the cases where the districts do stand together most closely. However, although at any particular age or grade one locality may seem to be considerably above another, still it is always true that at some other point it falls below. The averages for each district, on the other hand, become consistently higher with age increase and grade progress.

In our study, we have included schools as far west and as far east as possible, and very near the northern and southern boundary lines. (There is, in some cases, more distance between two rural districts than between any rural and urban location). We have found no clear cut differences between the various districts studied. We may conclude, then, that our tests do not tap information indigenous to one particular locality.

This is an extremely important point to ascertain. When, as is sometimes the case, tests are scaled and standardized on children in one locality only, it is doubtful if reliable group comparisons can be made with another locality. Here again the original norm material must be carefully scrutinized.

Differences Between One-Room and Union Schools.—A general description of the nature of one-room and union schools has been given above. The Census Bureau has estimated that at the end of 1922 there were 175,000 one-teacher and 13,000 consolidated schools in the United States. The consolidated schools are, of course, constantly on the increase.

All of our 638 high school subjects and 1170 of our grade school children were recruited from union schools. We have treated results from the two types of schools separately in all cases. Table XVII shows the age and grade averages for boys and girls in the union schools and the one-room schools. Scores from all union schools and all one-room schools respectively have been combined in this table.

TABLE XIII
 SCALED INFORMATION B
 AGE SCORES IN ONE-ROOM RURAL SCHOOLS

<i>School District A</i>				<i>School District B</i>				<i>School District C</i>			
<i>Age & Sex</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>		
9 <i>B</i>	9	10 56	4.20	6	11 34	2 92	10	11 80	5 30		
<i>G</i>	8	9 24	3 66	10	12 20	4 84	15	10 74	3 50		
10 <i>B</i>	17	11 58	3 06	17	12 30	3 94	33	12 10	4 52		
<i>G</i>	17	11 94	2 58	33	11 84	4 18	31	11 46	3 50		
11 <i>B</i>	28	14 50	3 54	24	12 66	3 74	25	14 76	4 00		
<i>G</i>	22	14 64	3 60	21	12 62	4 16	20	13 50	4 10		
12 <i>B</i>	31	13 26	4 24	33	15 54	4 06	32	14 44	4 04		
<i>G</i>	41	14 76	4 16	37	14 18	4 20	29	14 58	4 68		
13 <i>B</i>	43	16 22	3 34	25	16 68	3 32	31	15 32	4 60		
<i>G</i>	20	15 60	4 24	23	16 28	3 26	27	15 96	3 78		
14 <i>B</i>	23	18 48	2 78	29	16 94	4 12	29	15 90	3 82		
<i>G</i>	20	17 10	3 32	21	16 42	2 04	20	15 50	3 94		
15 <i>B</i>	24	17 34	3 30	13	17 16	3 46	22	15 72	4 80		
<i>G</i>	17	16 88	2 88	18	16 78	3 20	9	17 66	3 12		
16 <i>B</i>	5			8	17 50	4 10	12	14 84	4 36		
<i>G</i>	11	16 64	3 16	9	17 44	3 24	4				
<i>School District D</i>				<i>School District E</i>			<i>School District F</i>				
9 <i>B</i>	6	13 34	2 12	5			5	15 00	3.58		
<i>G</i>	10	12.20	3 70	10	12 40	2 54	13	11 00	2 94		
10 <i>B</i>	24	13.92	3 38	19	11 74	3 80	30	12 46	4.22		
<i>G</i>	18	13.56	4 52	30	11.06	3 16	46	11.78	3 54		
11 <i>B</i>	21	13 28	3 96	28	13 22	5 22	42	13 24	3 16		
<i>G</i>	21	14 80	4 62	27	13 74	3 60	56	13 42	4 32		
12 <i>B</i>	31	16 10	4 56	33	15 18	3 42	67	16 14	3 92		
<i>G</i>	22	14 64	4 20	40	14 24	4 56	60	14 36	4 46		
13 <i>B</i>	16	17 88	3 08	36	17 28	3 60	56	15 96	4 68		
<i>G</i>	12	16 34	3 08	36	14 56	3 78	51	15 32	4 20		
14 <i>B</i>	20	18 10	3 44	22	18 10	3 06	49	16 60	4 26		
<i>G</i>	17	18 42	2 56	18	15 00	4 16	37	16 24	3 62		
15 <i>B</i>	18	18 56	2 94	22	18 82	2 96	32	17.88	3 28		
<i>G</i>	10	17.20	4 60	15	17 26	3 34	14	16 86	3 24		
16 <i>B</i>	5	16 60	1 96	14	17.86	2 60	8	18 24	2 82		
<i>G</i>	2			12	16.00	3 16	8	15.76	3.86		
<i>School G*</i>											
<i>Age</i>											
10 <i>B</i>	12.10	4.12	11								
12 <i>B</i>	18.40	3 24	10								

*In this school, the numbers were so small that only at two ages was it possible to compute averages.

TABLE XIV
SCALED INFORMATION B
GRADE SCORES IN ONE-ROOM RURAL SCHOOLS

<i>School A</i>				<i>School B</i>			<i>School C</i>		
<i>Grade and Sex</i>	<i>No of Cases</i>	<i>Aver.</i>	<i>σ</i>	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>	<i>No of Cases</i>	<i>Cases</i>	<i>σ</i>
3 B	3								
G	3								
4 B	65	11 90	3 74	35	10 88	4 38	50	10 16	4 28
G	28	10 72	3 46	37	11 06	4 48	47	9 90	3 30
5 B	41	14 66	3 18	34	14 18	2 92	33	13 12	3 28
G	38	12 78	3 54	26	12 84	3 38	37	13 86	2 66
6 B	27	16 70	2 44	34	16 36	3 34	56	15 36	3 34
G	45	15 26	3 00	23	15 60	2 82	33	15 48	3 34
7 B	25	18 84	2 64	37	18 02	3 00	36	16 94	3 10
G	30	18 26	2 66	31	17 46	2 00	30	17 40	3 16
8 B	22	18 90	3 04	16	18 00	3 44	24	19 00	2 52
G	13	19 46	2 38	24	17 76	3 00	11	18 64	2 82
<i>School D</i>				<i>School E</i>			<i>School F</i>		
3 B	32	8 68	4 00	38	10 64	4 40	46	10 92	4 10
G	18	8 56	4 70	70	10 20	3 70	58	9 62	3 60
4 B	32	12 00	2 96	43	13 98	3 20	78	14 36	3 12
G	34	11 00	4 34	28	13 86	2 06	65	12 66	3 92
5 B	26	14 92	2 70	42	17 28	3 08	55	15 76	3 74
G	18	15 34	3 00	51	15 32	2 76	55	15 00	2 40
6 B	26	17 00	3 88	36	18 94	2 06	67	17 02	3 42
G	25	16 76	3 22	29	17 20	3 12	65	16 88	2 90
7 B	34	18 06	3 18	25	18 92	3 18	51	19 90	2 70
G	15	17 80	3 00	22	17 72	2 72	35	17 98	2 84
8 B	20	19 40	2 06						
G	18	17 22	2 82						
				<i>Gr.</i>	<i>School G*</i>				
				4 B	9 72	4 26	14		
				5 B	14 84	2 52	12		
				7 B	16 78	4 26	9		

*In this school, the numbers in each grade were so small that only at three grades was it possible to compute averages.

TABLE XV
SCALED INFORMATION B
AGE SCORES IN UNION RURAL SCHOOLS

<i>School A</i>				<i>School D</i>			<i>School E</i>		
<i>Age and Sex</i>	<i>No. of Cases</i>	<i>Aver.</i>	σ	<i>No. of Cases</i>	<i>Aver.</i>	σ	<i>No. of Cases</i>	<i>Aver</i>	σ
9 B	1			4			11	10 46	3 68
G	5			3			13	11 30	3 90
10 B	8	9 26	4 30	10	16 20	1 84	18	14 12	3 22
G	12	12 16	4 20	9	13 00	3 76	24	11 76	3 26
11 B	17	12 18	3 82	6			29	14 72	3 82
G	15	12 60	3 78	6			36	13 22	4 18
12 B	14	15 14	5 78	14	14 72	4 14	37	15 12	4 44
G	19	15 10	4 28	5			28	14 72	3 92
13 B	18	17 00	3 26	6			18	18 44	3 38
G	20	14 90	3 38	5			30	14 66	4 14
14 B	17	18 18	4 56	8	19 00	1 74	39	17 00	3 90
G	31	16 38	4 58	13	17 62	3 18	38	16 78	3 18
15 B	23	18 14	3 00	14	20 58	1 88	19	17 74	2 26
G	14	18 86	4 50	11	16 82	2 88	26	17 08	4 70
16 B	17	19 12	2 88	12	19 66	3 92	33	19 00	3 54
G	16	18 12	3 32	15	19 40	2 22	20	17 60	3 16
17 B	9	20 56	2 64	11	21 36	2 94	14	18 14	3 52
G	10	18 40	2 00	13	20 08	1 86	16	17 38	2 84
18 B	9	20 34	2 30	15	21 14	2 96	20	20 06	2 42
G	8	21 26	2 10	14	20 72	1 48	18	19 00	3 12
<i>School F</i>				<i>School G</i>			<i>School H*</i>		
8 B							19	5 42	3 46
G							28	4 58	3 44
9 B	3			1			34	6 24	4 20
G	9	12 56	4 20	3			40	7 90	4 34
10 B	12	11 84	3 32	5			42	9 72	5 18
G	30	12 54	2 01	10	12 60	3 56	46	8 86	4 14
11 B	22	14 28	3 80	18	15 66	4 10	29	12 72	3 96
G	29	14 38	4 20	8	16 00	3 00	40	14 04	3 84
12 B	24	16 74	4 44	24	15 92	3 40	39	12 94	5 78
G	43	16 40	4 02	8	15 50	3 58	53	14 70	4 94
13 B	30	17 40	3 60	20	18 70	3 96	52	16 24	3 92
G	32	16 68	3 24	18	15 88	3 34	46	16 44	3 06
14 B	39	19 16	3 32	24	19 08	2 98	42	16 48	3 78
G	49	17 82	3 30	14	15 56	4 68	27	18 18	3 18
15 B	39	19 30	2 38	27	19 30	4 00	37	16 40	5 50
G	41	18 32	2 92	20	18 00	3 54	50	17 64	3 50
16 B	23	18 92	2 92	11	21 54	3 42	35	18 36	2 28
G	30	18 40	3 24	13	18 54	4 38	22	17 82	3 06
17 B	34	20 58	3 56	6			16	20 38	2 20
G	35	20 14	2 04	16	20 38	3 06	16	18 38	2 98
18 B				9	21 88	1 92	5		
G				25	19 48	2 28	4		

TABLE XVI
 SCALED INFORMATION B
 GRADE SCORES IN UNION RURAL SCHOOLS

<i>School A</i>				<i>School D</i>			<i>School E</i>		
<i>Grade</i>	<i>No. of Cases</i>	<i>Aver.</i>	σ	<i>No. of Cases</i>	<i>Aver.</i>	σ	<i>No. of Cases</i>	<i>Aver.</i>	σ
4 <i>B</i>	24	10 26	3 78	9	12 56	3 10	42	10 86	3 74
<i>G</i>	29	9 68	3 42	15	12 46	2 96	45	10 46	3 60
5 <i>B</i>	17	13 82	3 42	17	15 24	3 36	35	14 88	2 52
<i>G</i>	26	13 62	3 34	6			36	12 72	2 54
6 <i>B</i>	20	16 30	4 02	8	17 00	3 00	36	16 06	2 68
<i>G</i>	24	16 42	3 02	4			26	15 54	2 80
7 <i>B</i>	13	17 00	2 34	7	18 14	1 82	30	17 40	3 62
<i>G</i>	14	15 86	2 60	13	17 92	2 44	37	15 64	3 38
8 <i>B</i>	17	18 88	3 02	4			41	18 96	2 74
<i>G</i>	18	17 44	2 76				36	17 62	3 00
9 <i>B</i>	23	19 60	2 72	15	21 40	1 50	18	20 12	2 24
<i>G</i>	18	19 22	3 12	20	27 70	2 92	32	18 68	2 42
10 <i>B</i>	9	19 22	1 98	17	20 30	3 06	16	20 76	2 98
<i>G</i>	5	19 80	2 04	14	18 86	2 44	15	17 40	3 36
11 <i>B</i>	9	21 88	2 34	6			8	19 50	1 94
<i>G</i>	15	20 74	1 90	20	20 90	1 48	7	19 56	1 40
12 <i>B</i>	6	21 34	1 38	17	21 82	2 38	12	21 50	1 20
<i>G</i>	7	21.58	2 06	13	19 78	1 68	12	19 84	3 20
<i>School F</i>				<i>School G</i>			<i>School H*</i>		
4 <i>B</i>	25	11 56	2 74	4			52	11 84	4 32
<i>G</i>	22	9 00	2 82	6			53	11 04	3 40
5 <i>B</i>	27	15 14	3 26	25	14 84	3 48	56	13 60	4 02
<i>G</i>	35	14 60	3 12	16	12 74	2 98	59	14 02	3 64
6 <i>B</i>	26	16 70	3 58	29	16 86	3 40	38	15 26	3 14
<i>G</i>	47	16 36	2 86	12	17 66	3 68	54	14 86	3 28
7 <i>B</i>	29	19 14	2 40	16	16 62	3 62	53	15 34	3 18
<i>G</i>	31	17 70	3 46	17	14 64	3 00	34	18 00	2 48
8 <i>B</i>	41	19 24	2 62	18	19 56	2 56	28	19 22	2 48
<i>G</i>	47	18 18	2 64	12	16 34	3 50	33	18 94	2 42
9 <i>B</i>	37	19 50	2 48	28	20 64	2 20	24	19 08	2 56
<i>G</i>	38	18 48	2 28	20	19 10	2 86	39	18 58	2 60
10 <i>B</i>	21	21 00	2 22	14	22 14	1 82	20	21 00	2 28
<i>G</i>	22	20 00	2 60	14	20 28	2 08	20	19 10	3 38
11 <i>B</i>	12	21 16	2 37	8	21 00	3 60 <i>Gr.</i>			
<i>G</i>	19	20 60	1 98	20	20 20	2 92 2 <i>B</i> 32		3 62	2.52
12 <i>B</i>	6	22 00	1 00	11	21 18	3 14 <i>G</i> 31		2 88	2 26
<i>G</i>	14	20 86	2 06	20	19 80	2 48			
							3 <i>B</i> 47	6 74	3 66
							<i>G</i> 53	6 24	3 42

*In this school, the different types of schools were not separated. Both one-room and union schools are represented.

TABLE XVII
COMPARISON OF ONE-ROOM AND UNION SCHOOLS

<i>Yrs. and Sex</i>	<i>One-Room</i>				<i>Union</i>			<i>One-Room</i>				<i>Union</i>		
	<i>No. of Cases</i>	<i>Aver.</i>	<i>σ</i>		<i>No.</i>	<i>Aver.</i>	<i>σ</i>	<i>Gr.</i>	<i>No.</i>	<i>Aver.</i>	<i>σ</i>	<i>No.</i>	<i>Aver.</i>	<i>σ</i>
8 <i>B</i>	22	9.54	3.74					3 <i>B</i>	38	8.42	3.86			
<i>G</i>	18	8.44	3.82					<i>G</i>	24	9.34	4.56			
9 <i>B</i>	43	11.98	4.56		20	10.80	3.84	4 <i>B</i>	280	11.02	4.08	104	11.14	4.12
<i>G</i>	66	11.30	3.70		33	11.48	3.92	<i>G</i>	278	10.32	3.78	117	10.32	3.44
10 <i>B</i>	151	12.38	4.02		53	13.50	4.04	5 <i>B</i>	267	14.24	3.12	121	14.84	3.18
<i>G</i>	179	11.86	3.72		85	12.32	3.78	<i>G</i>	219	13.36	3.60	119	13.56	3.08
11 <i>B</i>	175	13.46	4.14		92	14.32	4.02	6 <i>B</i>	245	16.32	3.46	119	16.50	3.36
<i>G</i>	172	13.78	4.30		94	13.90	4.14	<i>G</i>	235	15.42	2.90	113	16.24	3.08
12 <i>B</i>	237	15.40	4.16		113	15.58	4.46	7 <i>B</i>	244	17.76	3.18	95	17.80	3.16
<i>G</i>	232	14.46	4.36		103	15.64	4.12	<i>G</i>	207	17.42	2.90	112	16.36	3.38
13 <i>B</i>	212	16.32	4.04		92	17.94	3.56	8 <i>B</i>	164	19.16	2.92	121	19.16	2.70
<i>G</i>	178	15.50	3.88		105	15.64	3.62	<i>G</i>	129	18.02	2.82	113	17.70	2.86
14 <i>B</i>	176	17.14	3.84		127	18.38	3.68	9 <i>B</i>				121	20.00	2.40
<i>G</i>	135	16.44	3.52		145	17.00	3.80	<i>G</i>				128	18.60	2.68
15 <i>B</i>	133	17.52	3.76		122	18.98	3.00	10 <i>B</i>				77	20.80	2.64
<i>G</i>	87	17.06	3.32		112	17.90	3.78	<i>G</i>				70	19.26	2.84
16 <i>B</i>	54	17.04	3.56		96	19.38	3.42	11 <i>B</i>				43	21.10	2.66
<i>G</i>	48	16.62	3.32		94	18.36	3.32	<i>G</i>				81	20.38	2.14
17 <i>B</i>					74	20.02	3.34	12 <i>B</i>				52	21.58	2.16
<i>G</i>					90	19.48	2.64	<i>G</i>				66	20.22	2.46
18 <i>B</i>					53	20.92	2.54							
<i>G</i>					65	19.84	2.52							

Table XVIII presents the statistical reliability of the differences between these two types of schools for the various ages and grades. Examining the girls' results, we see that in the age-to-age comparison, the union schools make a consistently better showing though the difference is not always reliable. The average number of chances for a true difference is 80.7 in 100. Dividing the age groups into those below 12, and those 12 and over, our averages are 88.8 and 67 (chances in 100) respectively. This is what we should expect, because at the upper ages all the brighter children are necessarily in union schools since one-room schools have no grades above the 8th.

As far as the grades go, the honors are evenly divided. For the 4th grade—perfect equality; for the 5th and 6th grades, an average of 85.5 chances in 100 for a true difference in favor of the union schools. For the 7th and 8th grades, 91 chances in favor of the one-room schools.

Turning to a consideration of the boys' data, we find that there is a fairly reliable superiority of the one-room schools

TABLE XVIII
COMPARISON BETWEEN ONE-ROOM AND UNION SCHOOLS
Reliability of Differences—Girls

<i>Years</i>	<i>Actual Difference</i>	<i>σ Difference</i>	<i>Difference in σ units</i>	<i>Chances that true diff is above 0</i>	
(Favor of Union)					
9	18	83	22	59	in 100
10	46	49	94	83	in 100
11	12	54	22	59	in 100
12	1 18	45	2 60	99 7	in 100
13	14	45	31	62	in 100
14	56	44	1 27	89	in 100
15	84	51	1 6	94	in 100
16	1 74	59	2 9	99 8	in 100
<i>Grades</i>					
4	0			0	in 100
5	20	35	57	72	in 100
6	82	35	2 30	98 9	in 100
7	—1 06	28	—3 80	100	in 100 (1-rm)
8	— 32	35	— 91	82	in 100 (1-rm)
Reliability of Differences—Boys					
<i>Years</i>					
9	—1 18	1 11	—1 06	85	in 100 (1-rm)
10	1 12	64	1 8	96	in 100
11	84	53	1 6	94	in 100
12	18	50	36	64	in 100
13	1 62	47	3 45	100	in 100
14	1 24	40	3 1	100	in 100
15	1 46	42	3 5	100	in 100
16	2 34	60	3 9	100	in 100
<i>Grades</i>					
4	12	47	26	60	in 100
5	60	35	1 71	96	in 100
6	18	39	46	67	in 100
7	04	37	1	54	in 100
8	0			0	in 100

over the union schools at nine years. Otherwise the age results are consistently in favor of the union schools with an average of 93.4 chances. (The 9-year discrepancy may probably be explained as in Chapter III.) From ages 13-16 the difference is perfectly reliable. The normally advanced 13-year-old will, of course, be in high school.

Considering the grades, though, we have practically the same results as with the girls. In the 4th to 6th grades there is an average sigma difference (in favor of the union schools) of .81, which is, of course, not at all reliable. Above the 7th grade there is equality.

We can offer a plausible explanation for the sudden turn of the tables in favor of the one-room schools at the upper

grades. There is probably considerable selection at the upper ages. The one-room school children are more retarded for their age than the union school children, and with the fewer advantages and opportunities they have, it is not surprising that the dull ones drop out early. It has been shown that in New York three-fourths of the children in the one-room schools are in the first four grades (48). This is what we should expect. And if there *is* so much dropping out, is it not reasonable to suppose that there is in the topmost grades a veritable survival of the fittest?

We recently ran across the following surprisingly confirmatory statement by Foote (29), who made a study of 10,999 consolidated school pupils and 4653 one-teacher school pupils. "There is a pronounced belief among school people acquainted with the conditions that there is a higher relation between inefficiency and elimination in the one-teacher school than there is in the consolidated school. In other words, the one-teacher school retains a relatively larger number of its proficient pupils through the upper elementary grades and should, if all other factors were equal, show superior results of instruction in the latter grades."³² Foote's own results do not confirm this prophecy. Ours do, however, most markedly. The explanation lies in the fact that Foote was dealing with academic subjects whose results would clearly reflect the kind of teaching enjoyed. Our test would probably depend far more upon original innate ability than special teaching.

In regard to selection at the upper ages, the Assistant Superintendent of New York State writes as follows: "We are doing everything we can to encourage common school districts to send both their 7th and 8th grade pupils to a central rural or union free school whenever one is available. . . . It is true, however, that children are still prepared for high school in a large number of one-teacher schools. I think it is probably safe to say that when such children are still found in the one-room school, they represent two groups. Either they are the cream of the rural community because other pupils in the neighborhood of similar age and grade obtained their working papers as soon as they completed the 6th grade and reached the age of 14, and dropped out of school. Or else they are children who should actually be in high school,

³² 29, p. 346.

but are held back because they failed to pass some one of their Regents examinations for high school entrance.”³³

Turning to other studies, we find that we are in complete accord with the general conclusion that union schools exceed one-room schools in achievement.³⁴ Unlike most other investigators, however, we cannot show an entirely consistent or reliable difference except at those ages at which a fair comparison is impossible.

These comparatively small differences can probably be laid to the lesser opportunities in the one-room school. The union schools are far more elaborately equipped, both from the standpoint of health and educational facilities. The one-room pupils would doubtless be at a disadvantage in practically any field. No matter how much a test may tap “innate” intelligence, still the use of the tools at hand will contribute some small fraction to the final score. This would be sufficient to explain our differences.

If our differences are smaller than those of other investigators (as appears to be the case), it may be that it is because our test is more suited to both groups than those of other investigators. Standard tests, scaled on city children, have always been used and these may favor union school children. If so, we may here bring to bear, to a lesser degree, the argument used in relation to the comparison of city and rural children by such tests. This point will be elaborated in Chapter V.

SUMMARY

1. The rural schools were classified into one-room and union schools. The former represents the “typically” rural schools in outlying districts. The latter represents the consolidated schools serving five or more districts.

2. No reliable or consistent difference was found between the eight districts tested.

3. The union schools were found, in our study, to be fairly consistently (although not entirely reliably) superior to the one-room schools up to the 7th grade. Selection of cases above this grade in the one-room schools (and confirmatory evidence

³³ From letter from Miss H. H. Heyl to author.

³⁴ For a detailed account of the surveys dealing with rural schools the reader is referred to the bibliography (particularly to 30).

thereof) was offered as an explanation of the apparent change at the upper levels.

4. It is very important, since such differences exist, that material gathered for purposes of standardization, etc., should be carefully differentiated so that groups as heterogeneous as those represented above, should not be included in a haphazard manner.

CHAPTER V

RURAL AND URBAN DIFFERENCES ON INFORMATION TESTS

A AND B*

Nearly thirteen million people in the United States live in villages. Every eighth person in this country is a villager.³⁵ It is evident, therefore, that any blanket statement concerning the mentality of the rural population has a direct bearing on our population as a whole. We can hardly damn our country dwellers without incriminating ourselves rather deeply. The Census Bureau has estimated that with a total population of approximately 112 millions in 1924, there were a little less than "10 million children enrolled in rural schools in the open country and in villages and towns of 1000 population and under."³⁶

We are unable to cite any really comprehensive study of the mental differences between rural and urban children. The thoroughgoing research undertaken by Baldwin and Fillmore (22) has not yet been reported in full. There are, however, a fair number of partial studies which we shall summarize below. In many of these the investigation of rural mentality has been merely incidental to the main purpose.

Pintner, in 1917 (37) tested the school population (154) of a village of 913 inhabitants. Using his Mental Survey test with which he had previously tested a large number of urban children, he found that the median index of mentality of the villagers was 10% below the urban norm. He concludes that the more intelligent families are leaving the village community.

Pyle and Collings (40), in 1918, reported the results of giving the entire population of school children 8-18 (2000) of a Missouri county the Pyle tests. The rural boys had 72.7% of the urban boys' standing; the rural girls had 77.5% of the urban girls' standing.

* We wish to thank the following rural superintendents for their splendid cooperation with us in our work: Mrs. L. T. DeOlloqui, Carthage; Miss M. S. Rundall, Amenia; Mrs. R. E. Brown, Granville; Miss M. L. Isbell, Norwich; Miss M. L. Rodgers, Moravia; Miss M. G. Hoffman, Lewiston; Miss R. M. Libby, Canton; Mrs. E. D. Grubb, Potsdam; Miss R. S. Gandy, Dennisville, N. J.

³⁵ 25, p. 15.

³⁶ 26, p. 339.

Pressey, in 1920, attempted a comparison of rural and urban children with the Pressey Primer Scale. He believed that since this test does not "involve literacy, nor school training, children from country and city should meet the examination on equal terms."³⁷ Also the scale is given so early that home influences might be expected to have operated to a much less extent than later. Of 183 rural children of 6, 7 and 8 years, only 22% scored above the median for age, as determined from city children. A former study with the Pressey Group scale had showed that only 27% of the rural children tested above the urban median (39). "It would seem reasonable to conclude then," says Pressey, "that the differences found by both scales . . . between urban and rural children were real differences in intelligence."³⁸

Irion and Fisher (34), 1921, found that 361 rural school children (11-16 years of age) scored 10 points below the urban norm on the National Intelligence test.

Hinds, 1922, comparing 68 rural children in Texas with urban norms on the Otis test concludes "that the country child is lower in general mentality as measured by the group mental tests than the city child."³⁹

Book, in 1922, reported the results of testing 7748 Indiana High School seniors, 1194 of whom were recruited from rural schools. His battery of 10 tests contained a practical information test. The median score of the rural group was found to be below that of the urban group. However, Book adds the following: "Rural schools have a larger percentage of seniors making the most superior grade of intelligence, while the city schools have proportionally more seniors making a high average intelligence."⁴⁰

L. S. Hollingworth, in 1926, summed up her impressions as to rural mentality as follows: "As regards the comparative frequency of gifted children in urban and rural environments, we have not much information at present. Such data as bear on the subject indicate that we shall probably find a greater proportion of gifted in the cities except in districts so remote from means of transportation as to have precluded migration of intellectual deviates to the city. With the easy facilities for travel at present existing almost everywhere in

³⁷ 38, p. 92.

³⁸ *Ibid.*, p. 93.

³⁹ 32, p. 123.

⁴⁰ 24, p. 235.

the United States, it is not surprising that we find relatively unintellectual performance in mental tests among rural school children."⁴¹

Terman, in his "Genetic Studies of Genius" (3), strikes somewhat the same note. Half of the parents of his supernormal subjects "were born in cities of 10,000 population or over, and almost a quarter in cities or towns of 1000 to 10,000 leaving *only* (italics are ours) a quarter for rural districts and towns or villages of less than 1000."

Some studies in our field are contributed by English authors:

Bickersteth (23), in 1917, tested 1200 children in the elementary schools of the Yorkshire Dales and of Leeds. The former represents extremely isolated rural districts, the latter an urban settlement. On the whole, the Dales children were found better in memory tests and the Leeds children in reasoning tests.

Thompson (42), and later Duff (28), working with the Northumberland mental test, found that the rural districts gave results more than a year behind the large cities.

Turning now to rural surveys in the United States, eight major ones warrant mention:

In the Virginia Survey (50), in 1921, the rural children were found to be from 1-1½ years behind the city children in the school subjects measured.

In the same year, 10,000 children in the counties and cities of North Carolina (44), and in 1922, 16,700 pupils in Kentucky (46), were tested in school subjects. The results in both cases were similar to those cited above.

Works, in the New York Survey (48), in 1922, found that on the Sigma Silent Reading test the rural scores, grade for grade, were considerably below city norms. This was true of arithmetic, history and other academic subjects.

The Indiana Survey (45), in 1923, found the one-room 8th grade reading scores to be 1½ years below Indiana city schools and the union schools ½ year below. Similar results were found in other school subjects.

In his Texas survey, 1925, Works (49) measured rural and urban children with the National Intelligence test. The one-room schools were found to be about 1½ years behind the city schools, and the union schools about ½ year retarded.

⁴¹ 33, p. 58.

O'Shea (47), in the Mississippi survey, 1927, tested grades 1 and 2 with the Pintner Cunningham scale, grades 3-8 with the National Intelligence test, and grades 9-12 with the Terman Group test. A significant difference in favor of the urban children was found to obtain.

Myers found that the Pennsylvania rural schools (10,621 8th grade pupils), on the Otis Classification test, rated 11% below the standard (36).

It will readily be noted that the evidence presented so far is extremely one-sided. Everyone agrees in favoring the urban child as far as mentality and school achievement is concerned. It is true that we have omitted, for the time being, one or two studies to be discussed later, but the bulk of the experimental work has been reviewed. We may now turn to the results of our own study.

The T scaling and standardization of Scaled Information A (on urban children) has already been described. It seemed to provide a good basis for the comparative study of urban and rural children. Six hundred and ten children from 50 schools in one rural district were tested. The administration of the tests was conducted as before. The individual teachers gave the test according to specific instructions. This method is particularly appropriate in the case of rural chil-

TABLE XIX
COMPARISON OF RURAL (DISTRICT I) AND URBAN GROUPS ON
INFORMATION A

Years	Rural			Urban				
	No of Cases	Aver.	σ	No of Cases	Aver.	σ	Diff in σ Units (Favor Urban)	Chances that true diff. is above 0
9	42	39 1	8 4	371	42 2	9 3	2 2	98 6 in 100
10	65	42 7	8 1	768	43 6	9 3	8	79 in 100
11	74	44 9	8 8	883	47 5	9 3	2 5	99 4 in 100
12	103	46 5	9 9	886	50 1	10 0	3 4	100 in 100
13	105	50 1	10 2	952	54 7	10 1	4 4	100 in 100
14	90	54 7	9 1	838	57 4	9 3	2 7	99 7 in 100
15	86	56 8	11 0	614	59 4	9 1	2 1	98 in 100
16	45	59 3	10 1	530	61 7	8 8	1 5	93 in 100
<i>Grades</i>								
4	127	36 3	6 8	970	40 7	9 0	6 5	100 in 100
5	125	44 6	7 3	988	47 8	9 0	4 5	100 in 100
6	116	50 1	6 3	955	50 2	9 0	2	58 in 100
7	96	53 3	5 5	932	54 7	8 1	2 3	98 9 in 100
8	89	59 3	7 0	822	58 6	7 9	— 9	82 in 100 (favor rural)

dren. Pressey has pointed out the fact that these children are shy and not at ease with strangers. Names of the children and schools were omitted, so that there should be no impetus for cheating.

Table XIX shows the results in comparison with the urban figures. With the single exception of grade 8, the urban groups exceed the rural. The reliability of these differences is also contained in Table XIX. For the ages, there are 96.9 chances, on the average, for a true difference. The rural children are about a year retarded all the way along. Graph IV illustrates this clearly. The 13-year rural score is equal to the 12-year urban; the 16-year rural to the 15-year urban, etc. The grade differentiation is entirely reliable at grades 4 and 5. But above this the differences vary until at grade 8 the difference is in favor of the rural children. This is probably explainable on the basis of selection as we have already suggested in the sex comparisons.

On the whole, then, our results agree with those of previous investigators. The urban children are approximately one year in advance of the rural children.

Various explanations have been advanced to account for this deficiency. Pyle and Collings put forth the following possible reasons:

1. The city children are of better stock.
2. The environment of the city hastens development.
3. Better teachers and schools of the city give training that enables the children to understand better what is expected of them.⁴²

Pressey has commented on the fact that rural children are at a disadvantage through shyness, etc., when tests are administered by strangers.⁴³ O'Shea states that "the superiority of urban over rural pupils may be due to superior educational facilities in the way of more capable teachers, more extensive educational equipment and more time spent in school each year, as well as to superior native ability. It is probable that there is a selective process in operation . . . leaving the more capable stock to locate in the cities."⁴⁴

These quotations represent the general viewpoint of most of our authors. Environment in the form of poor opportuni-

⁴² 40, p. 538.

⁴³ 38, p. 96.

⁴⁴ 47, p. 226.

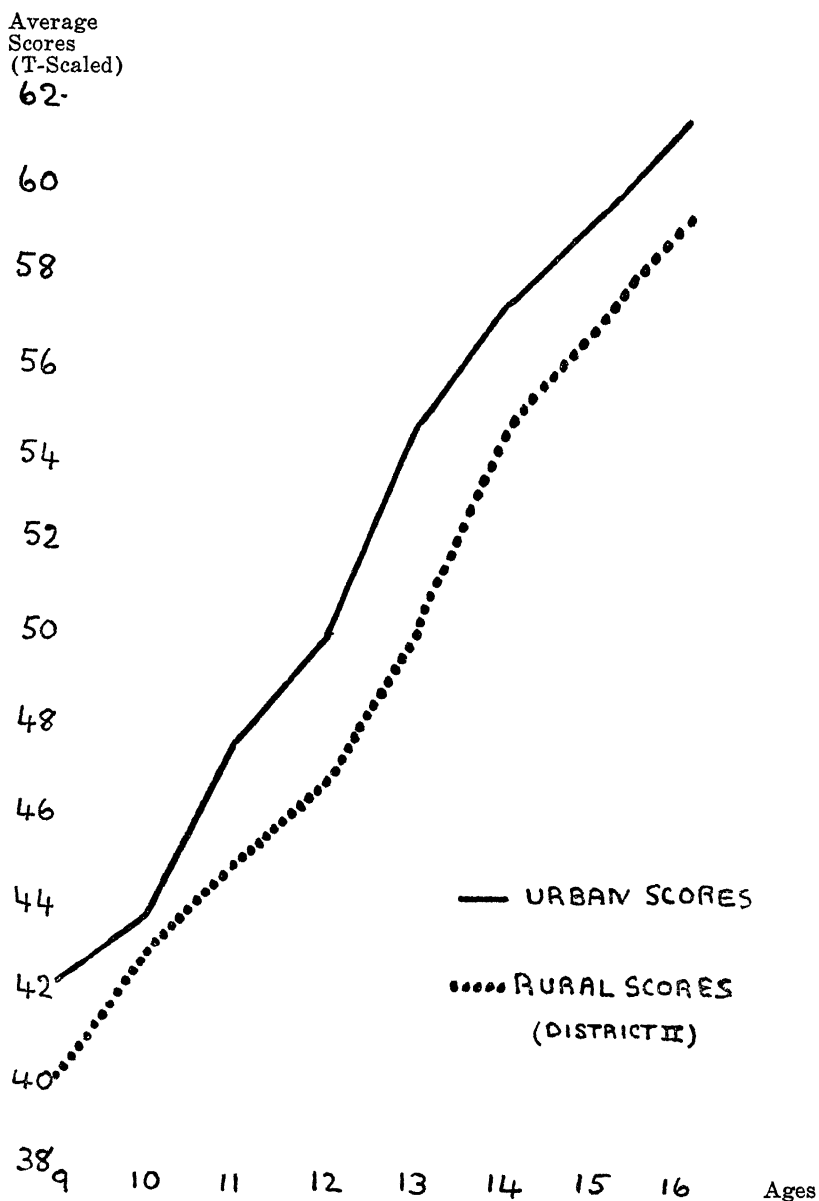


FIG. 4. Information A. Urban and Rural Scores

ties of the child, or heredity, since the better stock has left, are invoked to account for the inferiority of the rural child. For the most part this *inferiority* is taken for granted. There is still another possible explanation, however. Perhaps the tests do not fit the rural children as well as they do the urban. The country children may be *different*, not inferior. A sociologist has recently stated our viewpoint rather clearly. "The farmer is neither peculiar nor unique, and not even inferior; he is just different. "Did you," says Halstead, "ever try to drive 13 pigs out of a cornfield when they did not want to go?" The city man has had no such experience. The problems of the farmer are not those of the city man and consequently his stock of ideas is not the same. If a word reaction test were given to a group of farmers, in 99 cases out of 100 the word Chicago would bring forth the response, "Sears Roebuck and Co.," while this response would not be at all likely to appear in a test given to city dwellers. The apperception masses of these two groups differ greatly, and most naturally the responses are divergent."⁴⁵

Perhaps our explanation lies not in the superior environment or heredity of one group over the other, but in the tests used to compare them.

It is certainly true that the majority of our tests are standardized on white urban children. Even where state-wide surveys have been made, the tests used were originally scaled on city children. In the first place, large numbers are more readily available in city schools. In the second place, city children have been considered "representative." And since it is urban educators who, as a rule, have made the tests, this is, of course, readily understandable. The same thing is true of inter-racial and national differences. The *white* man's tests are used as a standard with disastrous results for the other peoples.

In racial psychology, this unfairness of the testing is being realized in many quarters, and attempts to form an adequate universal test are being made. A glance at almost any of our current psychological magazines, however, will serve to show that Italians, Indians, Japanese, etc., are still being tested with our group tests and conclusions drawn with or without reservations. We shall discuss the absurdity of this in greater detail in our next chapter. One quotation will suffice here:

⁴⁵ 43, pp.771-2.

"Some of the most commonplace experiences, for the average white," says Garth, "are lacking to the average Indian. For instance, the hogun and tepee have no 'chairs,' 'tables,' 'cellars'. . . . But a knowledge of such commonplace things and their significance in civilization is presupposed in those who are to pass satisfactorily the white man's intelligence tests."⁴⁶

It is unnecessary to labor this point. In such cases it is obvious. However, the same thing may be true in a much more subtle way of the differences between country and rural children. For some reason, our city tests may not fit our country cousins, so to speak.

So, as has been described, we attempted to make a test for rural children in exactly the same fashion as a test had been constructed for the urban group. This B test was then given

TABLE XX
GRADE DISTRIBUTION OF RURAL AND URBAN CASES—(Information B)

Grade and Sex	Rural			Total	Urban
	<i>1-Rm</i>	<i>Unon</i>	<i>Undif.</i>		
3 <i>B</i>	38	3	47	88	
3 <i>G</i>	24	33	53	110	
4 <i>B</i>	280	104	52	436	136
4 <i>G</i>	278	117	53	448	161
5 <i>B</i>	267	121	56	444	119
5 <i>G</i>	219	119	59	397	114
6 <i>B</i>	245	119	38	402	151
6 <i>G</i>	235	113	54	402	125
7 <i>B</i>	244	95	53	392	79
7 <i>G</i>	207	112	34	353	77
8 <i>B</i>	164	121	28	313	
8 <i>G</i>	129	113	33	275	
9 <i>B</i>		121	24	145	
9 <i>G</i>		128	39	167	
10 <i>B</i>		77	20	97	
10 <i>G</i>		70	20	90	
11 <i>B</i>		43	4	47	
11 <i>G</i>		81	1	82	
12 <i>B</i>		52	2	54	
12 <i>G</i>		66	4	70	
Total	2330	1808	674	4812	962

⁴⁶ 57, pp. 383-4.

to the original urban and rural districts and to seven other rural districts. The distribution of cases may be seen in Table XX.

Table XXI and Graph V show the comparison of urban and rural children on Test B. (Incidentally the sex difference is clearly illustrated.) One-room school children have been taken as the basis of comparison. Whereas, for conventional reliability $\frac{\text{Difference}}{\text{Sigma } \sigma}$ must equal 3, here the $\frac{\text{Difference}}{\text{Sigma } \sigma}$ ranges from 5.56 to 9.33.⁴⁷ The conditions found on Test A (illustrated in Graph IV) have been exactly reversed and very emphatically.

TABLE XXI
RURAL AND URBAN SCORES ON INFORMATION B
Comparison Between One-Room and Urban (3 Schools) Averages *

Grade and Sex	Rural		Urban			
	Aver.	σ	Aver.	σ	Diff. in σ Units	Chances that true diff. is above 0.
4 B	11 02	4 1	8 74	3 9	5 56	100 in 100
G	10 32	3 8	7 94	2 9	7 21	100 in 100
5 B	14 24	3 1	10 42	4 0	9 10	100 in 100
G	13 36	3 6	9 44	3 8	9 33	100 in 100
6 B	16 32	3 4	13 46	4 0	7 33	100 in 100
G	15 42	2 0	11 72	3 8	9 25	100 in 100
7 B	17 76	3 2	13 84	3 6	8 52	100 in 100
G	17 42	2 9	13 38	3 6	8 78	100 in 100

*Also see Table XXVII, Chapter VI.

A handful of other investigators have also been loath to conclude that urban children are superior to rural.

Gray and Marsden (1922), in England, tested children in the Yorkshire Dales and compared the results with those of town children. They conclude that "the country children examined are, as a group, more intelligent than *some* town classes we tested. . . . At least . . . our results do not support the assumption that the children in the country are much less intelligent than the children in the town."⁴⁸ The group studied was so small, however, that results can be considered only very tentative.

⁴⁷ The differences between one-room and union schools, elaborated in Chapter IV, pale into insignificance beside these results.

⁴⁸ *31*, p. 231.

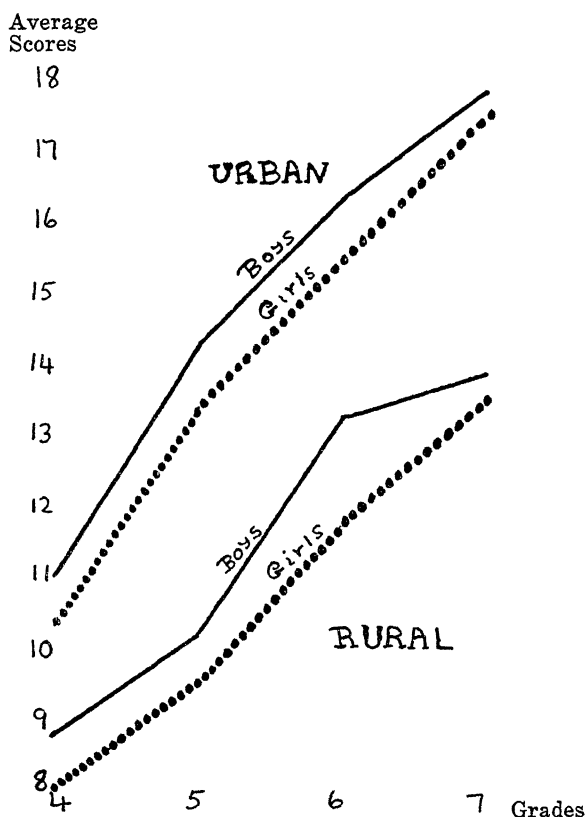


FIG. 5
Information B. Urban (Average 3 Schools)
and Rural (Average One-Room Schools)
Scores

The most important study has, unfortunately, to date been only briefly reported. At the American Psychological Association meeting, 1928, Baldwin and Fillmore reported as follows: "A study was made of all the children from birth to 16 years of age in four rural communities in Iowa in which children attend one-room schools or consolidated schools, and of a control group in an Iowa city with a population of 15,000."⁴⁹ "When compared in intelligence with children at large by means of established norms, and when matched with city children, the rural infants show no noticeable differences; the rural pre-school children show some inferiority at the upper

⁴⁹ 22, p. 185.

ages, and the rural school children show mental retardation that becomes increasingly apparent as they progress through school. . . . An analysis of the results of verbal intelligence tests reveals a striking difference in the language ability of rural children in contrast to city children. An analysis of the non-verbal reactions to the performance tests shows that while the rural children are handicapped by a slower rate of action, they show superiority on certain tests that probably relate to their experiences."

"These results raise the fundamental question: Are these inherent differences between the mental traits of rural children and of city children? Are the differences that have been found due to development and environment? Or are any of the present tests for measuring traits and their development adequate for rural children?"⁵⁰

We contend that they are not, and believe that the point has been at least partly proved by the demonstration of the fact that urban pupils do as poorly on rural-scaled tests as vice versa.

Several objections may be raised here. In the first place, have we perhaps so selected our rural questions that they favor our rural subjects unduly? It would, of course, be easy to pick questions that would be so specialized in character as to ensure the inferiority of any other group tested with them. We do not believe that this argument can be fairly launched against us. Our 80-question preliminary sheet for the rural test was chosen in precisely the same manner as the 80-question preliminary urban test, *i.e.*, from the questions submitted by the teachers of the children of the group. Any bias on the part of the author (herself distinctly an urban product) was in favor of the urban children, since with the rural tests all questions were eliminated which seemed to her to be at all specialized. In the final test, moreover, it must be remembered, the questions were not selected on the basis of ease of response for rural children, but ran from very easy to very difficult for that group. If, in the more or less mechanical selection of questions with the same number of sigma units of difference between them, questions particularly favoring the rural children were chosen is not this criticism equally true of our urban test and of practically all other tests? Our

⁵⁰ *Ibid.*, pp. 185-6.

B test is certainly no less fair to urban children than all other tests may be to rural children.

We have analyzed in detail the B papers of a number of rural and urban children (taken in exactly similar proportion from grades 4-7). On question 4 the urban children had a larger percentage of correct answers. (See Table XXII.) Questions 2, 13, 14, 15, 17, 18, 20, 21, 25 seem to be equal in difficulty for the two groups.⁵¹ On the remaining 15 questions, the average sigma difference in favor of the rural children is .53.

TABLE XXII
QUESTIONS FAVORING URBAN OR RURAL GROUPS (Information B)
Comparison Between School II and 3 Rural Districts

Question	σ Difference (<i>Favor of Rural</i>)
1	.48
2	.18
3	.47
4	— .57
5	.20
6	.49
7	.35
8	1.64
9	1.05
10	.47
11	.63
12	.47
13	0
14	— .07
15	.19
16	.21
17	— .19
18	.05
19	.58
20	— .02
21	— .06
22	.30
23	.40
24	.20
25	— 16

Considering that the sigma values run as high as 4.61, this difference is so comparatively small that we may say confidently that though, of course, these questions did favor the rural children, as the A questions favored the urban children, still they were not at all unfamiliar to the latter group or outside his scope of information.

We also attacked this problem from another angle. It will

⁵¹ In this study, a difference of less than .2 sigma has been discounted.

be remembered that nearly 50% of the questions in the original preliminary scale to the *urban* test were included in preliminary scale B. Studying these 36 overlapping questions (shown in Table XXIII), we find that 17 are passed equally by the rural and urban groups, 8 are to the favor of the urban group and 11 to the advantage of the rural group.

TABLE XXIII
OVERLAPPING QUESTIONS
Favoring Rural or Urban Children²

	σ Difference (Favoring Urban)
How many cents are there in a quarter?	— 22
What may we expect when we see heavy black clouds?	— 90
What is our national song?	20
How much does it cost to mail a letter to any city in the U S ?	— 14
Of what is butter made?	— 80
What holiday do we now celebrate that was first celebrated by the Pilgrims?	52
How many sides has a triangle?	.35
How old must you be before you can vote?	10
Of what is paper made?	16
How many states are there in the U. S ?	— 27
Name four different trees	06
Who was the first president of the U S ?	— 03
How many pints are there in a quart?	— 30
What do the stars in the American flag represent?	— 06
What is the capital of the U. S ?	— 12
What is the shortest month in the year?	.20
What is a submarine boat?	.35
Who is the Governor of your State?	— .07
What is the largest city in the U S ?	— .48
Who was the President of the U. S. during the World War?	67
Why should we kill flies?	— 49
Why is it dark at night?	.03
Name a country in Europe which is a republic.	13
How many weeks are there in a year?	— .02
Why don't we see the stars in the daytime?	— 38
About how often do we have a full moon?	— 06
Where does Congress meet?	09
What causes an eclipse of the sun?	.23
Why do we celebrate the 4th of July?	— 30
What artificial waterway connects the Atlantic with the Pacific?	— .17
What is steam?	— 02
What form of government have we in the U. S.?	— 08
What is the economic value of Alaska to the U. S ?	— 14
Of what is rubber made?	—1 16
Why is the moon light at night?	.46
Name two stones used for building purposes.	— 28

Leaving out one large difference (for the question "Of what is rubber made?") the average superiority for the urban group is .37 sigma, and for the rural group .44 sigma. Hardly a substantial difference! This would seem additional proof for our statement that the groups are different; not inferior or

superior. That there is a slight advantage, on the whole, for the rural group is interesting. These overlapping questions were all in the preliminary urban test, but necessarily of those not included in the final test. The first part of the preceding statement shows that we did not intentionally frame our test to fit one specific group. But, as the latter part of the sentence suggests, in the scaling the more urban questions apparently were automatically selected.

It may be well, at this point, to consider how we did get from 80 questions, which might presumably have fitted any group, an urban test A in the one case and a rural test B in the second. There are several steps in the testing process: (1) selecting the preliminary questions; (2) scaling the test from any group's percentage of correct answers on each question; (3) the standardization of the test. It is obvious that the last named process can not affect the test. We have shown that the selective process did not occur in the first step. By this we mean that either a rural or an urban test could have been scaled from the same original 80 questions. The second step must be the deciding factor. We select from the material tried out on the urban group, for example, questions evenly distributed throughout all the ranges of difficulty for that group. Apparently, however, this selection does not equally strike the various ranges of difficulty for another group. It may be that instead of one 1 question (*i.e.*, 1 in difficulty), one 2 question, etc., being included there may be three 4 questions for the second group. So, if the individual's range of ability is less than 4 he has a much lower score if he is from the second than if he is from the first group. It might, of course, happen that the test was easier for the second group; that the range of questions was skewed in the direction of the easy questions. This is not true, however, with either of our groups and is probably much less likely to be true.

We may say, then, as a general rule, that if, on a group homogeneous in respect to one quality, one scales a test according to the procedure outlined above, it will not equally tap all ranges of ability in a second group homogeneous in respect to a different quality.

The point has been raised whether this would be true of 8- and 10-year-old age groups, for example. This brings us into a somewhat different realm. A 10-year-old group is not really homogeneous in regard to a different quality than an 8-year-

old group. The first group is merely more mature, farther along the same path.

Even in age and grade groups, however, it would be possible, by scaling on one group alone, to produce a test not valid for other groups. In our scale, the preliminary questions were tried out on all ages and grades. Any question that seemed to single out one level was eliminated. For example, the question "Where are the Great Lakes?" elicited 50% correct answers from 7th graders and only 20% from 11th graders. This was obviously because this question is studied in the 7th grade and forgotten by the 12th. But this illustrates the fact that if one used only 7th grade children one might scale a test not entirely fair to even an older age group.

The two analyses quoted above also proved that we cannot tell *a priori* which questions are going to favor a certain group. A glance at Tables XXII and XXIII will bring this out clearly. It seems strange, for example, that the rural children do not have the advantage in "telling one way to find out the age of a tree," "giving one reason for the rotation of crops," "locating the Pole Star," etc.

To further substantiate this point (that questions favoring one group cannot be selected *a priori*), the overlapping questions plus a few others from Scale B were given to 14 rural superintendents, with the following directions: "Place a check mark beside each question which you think favors rural children (*i.e.*, in which rural children might be expected to excel), an equal sign where the chances of rural and urban children would be alike. If the question seems to favor the urban child, leave it unmarked."

The superintendents' answers were scored wrong *only* if they favored rural instead of urban or vice versa. Equal judgments were not recorded. The percentage of definite judgments made was computed. The average for our 14 rural superintendents was 82%. This seems high until we consider that with our scheme of marking, there was a 50% chance for a correct judgment. Under these conditions, such an average from a group of experts in the field is significant. If they can actually designate as distinctly rural what is apparently distinctly urban (or vice versa) on even two questions out of the 25, we certainly cannot safely decide *a priori* that any test favors or disfavors or is equally fair to any group or groups.

The same test was tried out on the Graduate Seminar in

Psychology at Columbia University. The 16 papers were recruited from members of the staff and graduate students. This might be considered, I think, a distinctly urban group. The average per cent correct was 83. This is practically identical with the findings reported above and is to be interpreted similarly.

Reisner, in a study of 8th grade pupils in Pennsylvania, conducted a somewhat similar investigation. A number of people acquainted with testing and rural and urban life were asked to classify the questions on Otis Self Administering—Intermediate Form B, as to whether they were strictly urban, strictly rural or intermediate. This brought out several interesting results. (1) There was a great difference of opinion among the judges themselves. (2) Questions might be correct from a rural standpoint but not included in the urban key, *e.g.*, "What is the most important reason that bright lights are placed in front of the theatres? so that people can see where they are; to attract attention and look inviting; so that people can see the advertisements better; electricity is furnished to theaters cheaply; to help light up the streets." Probably the first answer would be most reasonable in places where it is hardly necessary to present a lure to jaded tastes. (3) The urban children did better even on many questions which were classified as rural, especially where word meanings were involved. In computation, however, the rural group excelled.

We cannot subscribe to Reisner's conclusion: "This . . . brings out the fact that urban pupils score higher than rural pupils on the questions classed as strictly rural as well as on questions classed as strictly urban."⁵² Having obtained five judges who, with several differences of opinion, rated the questions, Reisner then concludes that certain questions *are* rural, etc. This is very fallacious reasoning. That it is possible to select questions which favor rural children is shown by our study. It is not wise to assume that the objects around one are necessarily familiar.

Both our own and Reisner's study seem to point conclusively to the fact that *a priori* judgments have little validity and are a very frail structure on which to build reliable comparisons.

In the second place we may be criticized for any general conclusions we may make since we have, unlike most of the

⁵² 41, p. 24. •

other investigators, used only an information test and not an intelligence test. We did this advisedly. We believe that it is the information difference that is so very important. A standard intelligence test is such a composite that one's total score doesn't tell a very complete story. But is it not true that information figures largely in practically each one of the tests? Some of the group tests contain sections avowedly informational, but such tests as analogy tests, definitions, etc., presuppose a certain amount of information. One may reason remarkably clearly, but the ease with which one completes

A policeman is to a burglar as a cat is to a . . .

is a function of one's familiarity with the duties of a policeman and the vocation of a burglar. It may be argued, of course, that the information required is so slight as to be the property of all concerned. This, however, is not true. At the higher levels the information required is almost prohibitive, as for example, in the vocabulary test of the Thorndike C.A.V.D. test, *e.g.*

Sub series Q. (To supply word in line meaning same as first word.)

21 radial	light, agitator, straight line, root, ray
22 sequestrate	follow, petition, horseman, confiscate, redwood
30 auricular	golden, heard, jointed, distinct, clear.

It may be objected that we are confusing language ability and information, but is the difference, after all, very real? If you've never seen or heard of an orange, you can't define it. And, moreover, if your way of living does not lead to the absorption of literature containing such words as "radial" or "auricular," you won't succeed well in this test. On the whole, of course, it is true that intelligence may be gauged by one's ability in such language tasks, but comparisons can be made only when the testees' environmental opportunities have been equal. As Thorndike himself phrases it: "The problem of analyzing a person's intellectual ability into the amount due to nature and the amount due to nurture is unsolved. No task or test has been proved to be a measure of the former alone. The wisest procedure at present is to equalize environmental forces by a wide variety of data with which all individuals have had adequate experience and to make as correct allowance as we can for what we cannot equalize."⁵³

⁵³ 4, p. 462.

But the selection of such tasks (*i.e.*, those with which all have had equal experience) cannot be made *a priori*. This has surely been abundantly demonstrated in the preceding section.

Thorndike has lately supported the hypothesis that "the higher forms of intellectual operation are identical with mere association or connection forming, depending upon the same sort of physical connections but requiring many more of them."⁵⁴ He submits this hypothesis "to an almost crucial test by determining the correlations within the upper half of intellectual operations with those in the lower half and those between the upper and lower halves." As a measure of the "higher" he used sentence completion, arithmetic problems and analogy tests. As means of more purely "association," vocabulary tests, routine and informational arithmetic and information tests. The "higher" abilities were found to correlate as closely with the associative as *inter se*, and vice versa. "These facts . . . prove that mere association and the higher abilities have in the main the same cause. Almost all of what is common to the one sort is common to the other."⁵⁵

To those, therefore, who would criticize us for making any general conclusions from results on information tests or allying our results with the broader studies of other investigators, we may say (1) that intelligence tests are, in some measure at least, information tests; and (2) that the same abilities are, to some degree, tested by mere information tests as by a more complex battery.

When we come to analyze the results of some of our authors who have given details of their work, we are driven still more firmly to the hypothesis advanced heretofore. In those cases where investigators have carefully analyzed their data they agree with us, in the main, in attributing the obtained differences to reasons other than to the innate inferiority of one group.

Pyle and Collings, (40), for example, find that the rural children approach the urban children more on the non-linguistic tests.

Chapman and Eby, comparing 15 one-room rural schools in Ohio with a large city school, conclude that "the superiority of the city school children over the one-room rural school children varied approximately in direct proportion as the de-

⁵⁴ *Ibid.*, p. 422.

⁵⁵ *Ibid.*, p. 430.

mands made by the test called for special school instruction as opposed to general powers which the school can do little to make or mar."⁵⁶

In several studies, as for example, the New York school survey (48), the rural children have been found to be equal to the urban subjects in arithmetic even though markedly deficient in other studies.

In the Bickersteth study, the Leeds (urban) children were inferior to the Dales (rural) children in memory work and not consistently better in cross-out tests, etc., but superior in reasoning (as measured by the Burt analogy test; examples of which are policeman : burglar : : cat : writing : typewriter : : voice :).

Bickersteth appends the following significant note: "If the superiority of the Dales children in the memory test were indeed a racial characteristic, we should expect to find it unchanged in the Leeds children of Dales ancestry, but of 31 such children, all but five were considerably below the average for their respective ages in the Dales group, but the same children were above the Dales average in the reasoning test."⁵⁷

Lehman and Witty, investigating the play activities of rural and urban children, find that their recreations are very different and believe that these differences are "directly traceable to environmental opportunities." The possible implications of their study they state as follows: "Numerous investigators have yielded data which show that rural children are somewhat below city children in mental age. . . . It is . . . plausible to assume that the lower mental age ratings of the rural children are . . . a result of the situation revealed by the above data. In administering mental tests, it is assumed that the individuals tested have similar environmental backgrounds, equal opportunities for acquiring information, etc. . . . It is evident . . . that the rural and the city children do not have the same social contacts, . . . certain it is that the environment of the town and country children are quite different and these environmental differences *may have* an influence upon the mental age ratings of the two groups of children."⁵⁸

Baldwin and Fillmore, to whose study we have already re-

⁵⁶ 27, p. 644.

⁵⁷ 23, pp. 66-7.

⁵⁸ 35, pp. 124-5.

ferred in some detail, found that whereas rural and urban infants and younger pre-school children were alike in mentality, the divergence between the two groups increased with age. They question whether the differences are inherent or due to development and environment, or whether we have any tests equally suitable for the measurement of the two groups. It must be admitted, of course, at the outset, that if the rural group *were* inferior, the difference would become more marked with age. Such a curve does represent our feeble-minded norms in divergence from the average. However, is not the following also a possible explanation of Baldwin and Fillmore's results? In infancy tests we are compelled to use more or less universal material. As the children grow older the tests become more and more linguistic and necessarily more and more tinged with environmental verbiage. If the thesis that we have set forth in this chapter is correct, this would also account for the growing inferiority of the rural group. And, with this explanation, certainly no more hypothetical than any other, Baldwin and Fillmore's results neatly dovetail with our own.

SUMMARY

1. The importance of data about rural children is evident when we realize that in 1924 there were 10 million children enrolled in rural schools.

2. The bulk of evidence from previous studies points to the conclusion that rural children are inferior to urban children mentally and scholastically.

3. We have offered the hypothesis that this difference is due not to any innate intellectual difference between the two groups, but to the tools used in measuring them.

4. To obviate this difficulty our Information Test B was scaled on rural children just as our Test A was scaled on urban children. Each test was administered to both groups of children.

5. On Information A the rural group was found to be about a year retarded in comparison with the urban group.

6. On Information B, the situation was entirely reversed. A perfectly reliable difference $\left(\frac{\text{Difference}}{\text{Sigma "a"}} \right)$ ranging from 5.56 to 9.33) was found between the performances of the urban and rural groups.

7. A handful of other investigators have come to conclusions that may be compared with ours—notably Baldwin and Fillmore in their very comprehensive study.

8. Analysis of urban and rural answers on Test B, and of overlapping questions in preliminary tests A and B, points to the conclusion that Test B is no more specialized in favor of rural children than Test A (or any standard test) is specialized in favor of the urban children.

9. From the same analysis we produced evidence that questions “fair” to a certain group cannot be selected *a priori*. This was also affirmed by submitting our questions to 14 rural superintendents who, despite their unusually rich experience, were unable to designate correctly (in a fairly large percentage of the cases) which questions favored the rural children.

10. An analysis of current standard group tests shows that a large part of the material required is informational in character. So, our results may be said to have some application outside the narrow sphere of individual information tests.

11. Through the entire rather involved chapter, we have attempted to follow the thread of this inquiry: “Are the mental differences found between urban and rural groups a function of an innate intellectual difference or of the tools of measurement?” Our results seem to point to the second explanation as possible and indeed probable.

CHAPTER VI

THE NATIONALITY OF THE SUBJECTS AND THE CORRESPONDING SCORES

The question of racial and national differences has long been a fascinating one, especially in this country which draws its peoples from such varied sources. The results of the Army examinations roused fear in many quarters, and gave rise to such statements as the following: "We are being swamped with the offscourings of Europe. . . . We have no place in this country for 'the man with the hoe' stained with the earth he digs and guided by a mind scarcely superior to the ox whose brother he is."⁵⁹ The influence of such a view (*i.e.*, of the native inferiority of other races than our own) on our social and economic policies is hard to estimate.

In the next few pages the outstanding studies on the various groups of foreign born in America have been sketchily summarized so that we may grasp, if possible, the general trend in this field. In examining these studies, we hope the reader will keep in mind the viewpoint of this paper, *i.e.*, that the tools of measurement must be evaluated as well as the intelligence of the subjects tested by them.

NEGRO STUDIES

A very short time ago, the current findings as to the racial inferiority of the negroes could be summed up as follows: "On the whole, there are found to be distinct differences between the two races considered, both as to intellectual ability, as measured by the tests used, and in school achievement. . . . It is shown that overlapping in intelligence is to the amount of 15 and 25% of the colored race who reach or exceed the median of the white."⁶⁰

Now, no such simple statement would be adequate or acceptable. In articles referred to in the bibliography, Peterson (65), Wells (70), and Sunne (67) have stressed the importance of environment and training and the impossibility of comparing the groups when, for generations, these factors have not been equal. Sunne and Davis (56) and Pressey and

⁵⁹ 68, p. 611.

⁶⁰ 71, p. 85.

Shively (2) have pointed out the probable unfairness to the negro group of test material now being used. Davis has shown the importance of equalizing educational opportunities. Two recent studies of Klineberg and Herskovitz may be briefly quoted here to show how far the pendulum has swung. Klineberg, 1927, used performance tests with white, indian and negro children. His results seem to show that the speed factor in the performance is largely environmental, not racial.⁶¹

Herskovitz measured with the Thorndike College Entrance examination 539 adult male negroes divided into eight classes, according to white characteristics. "The hypothesis of less negro intelligence and racial efficiency," says Herskovitz, "when compared to whites, which has been generally accepted from results in psychological tests, must be further tested by the acceptability of its logical corollaries." One of the chief of these, he believes, is that in mixture, those individuals having the most white blood should be superior to those having more negro blood. The correlations of the psychological with such physical characters as width of nose, thickness of lips, etc., were found to be entirely insignificant. Herskovitz discounts Ferguson's results on the ground that the discrimination within the negro group against those individuals showing such negroid traits as dark skin color, would cause differences in social environment which would affect the mean standing of groups selected on the basis of these traits."⁶² He concludes, in the light of his findings that "the basic hypothesis of white superiority in general social efficiency and innate intelligence is to be gravely doubted."⁶²

ORIENTAL STUDIES

All the comparative studies of orientals agree that the Chinese and Japanese children are not, on the whole, inferior to American children except where obviously handicapped by language difficulties. Particularly interesting to us are the results of Wang who, in 1926, tested Chinese and American college students (paired) with a series of tests. He found the Chinese decidedly superior to the Americans on number series tests, but decidedly inferior on *general information* which most involves language difficulty and knowledge of American life and customs⁶³.

⁶¹ 68.

⁶² 59, p. 42.

⁶³ 69, p. 104. Italics are ours.

INDIAN STUDIES

The bulk of the studies of Indian children has been contributed by Garth. He concludes that although on our tests the Indians score somewhat below the white, still "because of differences in social status and temperament we cannot conclude that our results are true and final measures of Indian children."⁶⁴ This is made more emphatic by the previous detailed discussion of the difficulties involved in giving our tests to Indians who lack our most commonplace experiences.

Helmer, in an unpublished thesis, reports an attempt to frame a test made up from experiences common to three Indian tribes. After trying out this test on both whites and Indians, she finds that those tests best adapted to the Indians are least adapted to the white, and concludes that "it would be difficult to make tests which would fit both the Indians and whites equally well."⁶⁵

Klineberg, 1928, finds with Indian children much the same results as those reported above with negro children, *i.e.*, "There is evidence that the superiority of white over Indian and negro children in performance tests is largely if not entirely a superiority in scores for *time*. There is no superiority; and in some cases an inferiority in the scores for *accuracy* of performance."⁶⁶

COMPARATIVE STUDIES ON OTHER NON-NATIVE GROUPS IN AMERICA

The question of whether verbal material is an important factor in our testing of foreigners, has been emphasized in many articles. There is some reason to believe that Italians rank below Americans on all our tests. As far as the other groups go, however, the two following quotations fairly sum up the trend of current opinion based on experimental evidence.

Hirsch, 1925, comparing his results on 11 groups with those of Brigham, says: "We can fairly and safely state that from 30% to 40% of the mental differences between the English and the combined non-English speaking groups as tested in the Army were due to a language handicap."⁶⁷

⁶⁴ 57, p. 389.

⁶⁵ 58, p. 76.

⁶⁶ 63, p. 107.

⁶⁷ 60, p. 341.

Kirkpatrick states that "there is no denying the fact that the superiority of the Americans is chiefly on the verbal tests."⁶⁸

There is some evidence to the effect that even where presumably "non-verbal" tests are used, the foreigners have still some language handicap.

Jones gave the Myers mental measure (presumably non-verbal) and two verbal tests to children of native and foreign parentage. The results seemed to point to the following conclusion: "The fact that a test is composed exclusively of non-verbal material is, therefore, no proof that it has merit as a non-verbal test."⁶⁹

Koch and Simmons, using the Meyers Pantomime and National Intelligence test with native and foreign children, make this statement: "While the pantomime test reduces to a minimum the language handicap of the foreign group, it is not to be assumed that the test permits all of our subjects to operate under equally favorable conditions. It is questionable, for instance, whether our groups have had equal opportunity to familiarize themselves with many of the concepts included in its content."⁷⁰

We wish to draw attention to the last sentence. This is precisely what we have inferred from the use of our tests with urban and rural, native and foreign-born children.

All in all, we may perhaps subscribe, at least to some degree, to the recent statement of Dale Yoder on the present status of the question of racial differences. "It may be correctly concluded that the consensus of competent scientific thought, contemplating the inability of mental testers to define intelligence, the inadequacy of all attempts to take such factors as education, social status, and language, into proper consideration and the deficiencies of testing conditions, finds no proof of racial inferiority or superiority and eliminates the usual methods of determining such standing from the field of scientific usefulness."⁷¹

We may safely say that there is a growing conservatism as to the conclusions we can draw, and more and more a dis-

⁶⁸ 62, p. 90.

⁶⁹ 61, p. 207.

⁷⁰ 64, p. 35.

⁷¹ 72, p. 470.

position to criticize our tools rather than our subjects in the case of our foreign-born.

However, whether or not experiments show conclusively that it is the language or special informational character of our tests that penalizes the foreigner, at any rate it will be universally conceded that the performance of foreigner and native on these tests is *different*. So it is extremely important to know whom one is testing. We have already commented on the use of deplorably small numbers in establishing norms. When we add to this the fact that the children in one school may be chiefly Italian while the children in the second may be Finn or Russian or fifth-generation Americans, it is obvious that comparisons are meaningless.

Armstrong's study (55) of rural and urban children is interesting in this respect. She tested 115 rural children in Katonah, N. Y., and 328 urban children in New York city. Contrary to usual results, the rural children, on both verbal and non-verbal tests, ranked higher. Analysis showed, how-

TABLE XXIV
PARENTAGE OF SUBJECTS—Information B

Subj. Group	No. of Cases	S. and 2 Parents Amer. Born	S and 1 Parent Amer. Born	S only Amer. Born	S. and 2 Pars. Foreign Born	No. of Countries Represented
<i>Rural</i>						
		%	%	%	%	
Dist. A	436	84	11	2	3	4 60% Canadian
Dist. B	262	77	6	16	1	11 46% Swedish
Dist. C	370	86	5	9	0	9 21% Austrian
						21% Czech
Dist. D	519	88	5	6	1	10 32% Austrian
Dist. E	561	55	19	20	6	13 23% Canadian
Dist. F	1026	72	15	11	2	15 22% Polish
Dist. G	338	94	2	3	1	5.
Dist. H	680	80	8	10	2	14. 25% Russian
						19% Italian
Average (weighted)		77	13	7	3	
<i>Urban</i>						
Sch. 1	283	13	16	65	6	14 39% Italian
						45% Finn
Sch. 2	247	73	12	9	6	10 21% Irish
						25% Canadian
Sch. 3	342	19	14	59	8	14. 63% Italian
						13% Russian
Average (weighted)	.	32	14	47	7	

ever, that the rural group was entirely American (for several generations) while the urban group was very heterogeneous nationally. When the latter was analyzed as to parentage it was found that differences were a function of the number of generations in this country.

We have analyzed our groups according to the birthplace of the parents. These data (in Table XXIV) are presented for

TABLE XXV
Urban Norms—Information B

<i>School 1</i>			<i>School 2</i>			<i>School 3</i>		
<i>Yrs and Sex</i>	<i>No of Cases</i>	<i>Aver.</i>	σ	<i>No. of Cases</i>	<i>Aver.</i>	σ	<i>No of Cases</i>	<i>Aver. σ</i>
9 <i>B</i>				18	10 12	4 24		
<i>G</i>				11	8 82	4 04		
10 <i>B</i>	30	8 80	3 32	34	11 42	3 82	38	8 74 3 98
<i>G</i>	49	8 56	2 64	30	9 26	3 00	35	8 54 3 92
11 <i>B</i>	35	11 00	4 22	25	13 32	4 30	44	9 82 4 18
<i>G</i>	56	9 74	4 00	41	12 08	4 28	36	8 06 3 72
12 <i>B</i>	47	12 36	4 08	50	15 04	3 48	46	11 66 5 16
<i>G</i>	36	11 22	4 08	42	13 80	3 82	37	10 08 3 82
<i>Gr. and Sex</i>								
4 <i>B</i>	48	8 62	3 38	48	10 38	4 44	40	6.90 2 82
<i>G</i>	68	7 76	2 56	47	8 92	3 00	46	7 24 3 12
5 <i>B</i>	31	10 22	3 76	31	12 42	3 30	57	9 42 4 12
<i>G</i>	44	9 90	3 64	23	10 14	3 16	47	8 66 3 88
6 <i>B</i>	46	13 40	2 60	53	15 26	3 10	52	11 66 4 88
<i>G</i>	43	10 82	3 30	47	13 46	3 82	35	10 48 3 46
7 <i>B</i>	25	14 36	3 28	13	16 54	2 38	41	12 66 3 72
<i>G</i>	19	14 16	3 52	24	14 84	3 26	34	11 94 3 40
Average of 3 Urban Schools								
<i>Yrs. and Sex</i>					<i>Gr. and Sex</i>			
9 <i>B</i>	32	9 26	3 96		4 <i>B</i>	136	8 74	3 90
<i>G</i>	26	8 08	3 52		<i>G</i>	161	7 94	2 94
10 <i>B</i>	102	9 62	3 94		5 <i>B</i>	119	10 42	4 02
<i>G</i>	114	8 74	3 18		<i>G</i>	114	9 44	3 76
11 <i>B</i>	104	11 06	4 44		6 <i>B</i>	151	13 46	3 98
<i>G</i>	133	9 96	4.24		<i>G</i>	125	11 72	3 80
12 <i>B</i>	143	13.08	4 52		7 <i>B</i>	79	13 84	3 62
<i>G</i>	115	11 80	4 00		<i>G</i>	77	13 38	3 62

both rural and urban groups on Test B. It will be seen at a glance that the rural districts are chiefly American. Only 10% have both parents of the subjects foreign-born, and a considerable percentage of these are English speaking. The three large urban schools, however, are strikingly different in make-up. Schools I and III have about 70% of foreign-born parentage (on both sides). School II is closely similar to our rural districts with 73% (as compared to 77%) with both parents American born. Fifty per cent of the foreign-born parents, moreover, are English speaking (*i.e.*, Irish and Canadian). Of the foreign-born in District A (which received both A and B tests) 60% are also Canadian.

Average
Scores

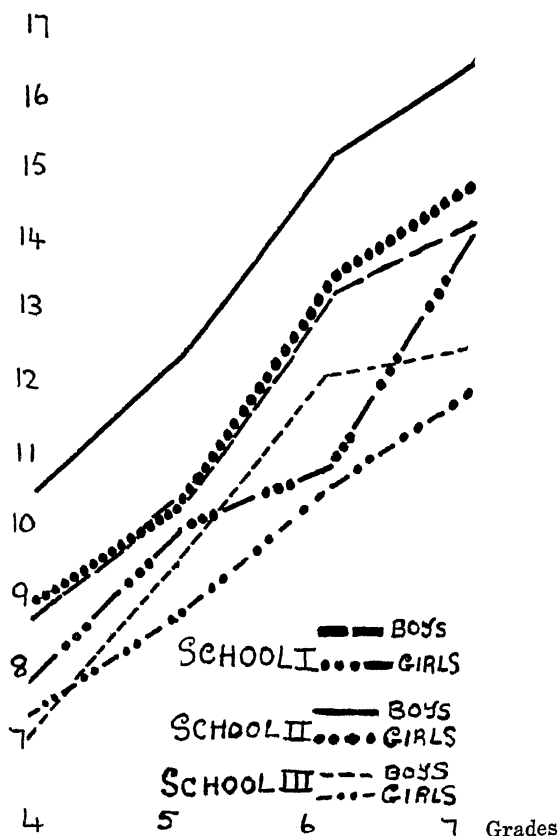


FIG. 6. Information B. Urban Scores

70 INVESTIGATION INTO THE VALIDITY OF NORMS WITH

Table XXV and Graph VI show the scores of the three schools on Test B. Table XXVI shows the reliability of the differences between these schools. There is a very significant difference between Schools II and III, and a consistent though not entirely reliable difference between Schools I and II, and Schools I and III.

TABLE XXVI
COMPARISON BETWEEN THREE URBAN SCHOOLS

Reliability of Differences—Information B									
School 1 and School 2					School 2 and School 3				
<i>Yrs. Sex</i>	<i>Actual Dif. (Fav. Sch. 2)</i>	<i>σ Dif.</i>	<i>Dif. in σ Units</i>	<i>Chances True Dif. Above 0 (in 100)</i>	<i>Actual Dif. (Fav. Sch. 2)</i>	<i>σ Dif.</i>	<i>Dif. in σ Units</i>	<i>Chances True Dif. Above 0</i>	
10 <i>B</i>	2 62	.90	2 91	99 8	2 68	93	2 88	99 8	
<i>G</i>	70	66	1 06	85	72	.86	.84	80	
11 <i>B</i>	2 32	1 12	2 07	98	3 50	1 07	3 27	100	
<i>G</i>	2 34	85	2 75	99 7	4 02	91	4 42	100	
12 <i>B</i>	2 68	.77	3 48	100	3 38	91	3 71	100	
<i>G</i>	2 58	.90	2 87	99 8	3 72	87	4 28	100	
<i>Gr and Sex</i>									
4 <i>B</i>	1 76	.81	2 17	98 6	3 48	78	4 46	100	
<i>G</i>	1.16	.54	2 14	98	1 68	.63	2 67	99.7	
5 <i>B</i>	2.20	.90	2 44	99 2	3 00	.81	3 71	100	
<i>G</i>	.24	86	.28	61	1 48	.87	1.70	96	
6 <i>B</i>	1 86	.57	3 26	100	3 60	.80	4 50	100	
<i>G</i>	2 64	.75	3.52	100	2 98	.81	3 68	100	
7 <i>B</i>	2 18	.84	2 60	99 5	3 88	.88	4 41	100	
<i>G</i>	.68	1.06	.64	74	2.90	.89	3 26	100	
School 1 and School 3					School 1 and School 3				
<i>Yrs. Sex</i>	<i>Actual Dif. (Fav. Sch. 1)</i>	<i>σ Dif.</i>	<i>Dif. in σ Units</i>	<i>Chances True Dif. Above 0 (in 100)</i>	<i>Gr. and Sex</i>	<i>Actual Dif. (Fav. Sch. 1)</i>	<i>σ Dif.</i>	<i>Dif. in σ Units</i>	<i>Chances True Dif. Above 0</i>
10 <i>B</i>	06	.89	07	52	4 <i>B</i>	1 72	66	2 58	99.5
<i>G</i>	.02	.76	.03	51	<i>G</i>	52	.56	.93	82 5
11 <i>B</i>	1.18	.96	1 23	88 5	5 <i>B</i>	80	.87	92	82
<i>G</i>	1 68	.81	2 07	98	<i>G</i>	1 24	79	1 56	93 5
12 <i>B</i>	70	97	.72	76	6 <i>B</i>	1 74	.77	2 26	98 7
<i>G</i>	1 14	93	1 23	88 5	<i>G</i>	34	.77	.44	67
					7 <i>B</i>	1 70	88	1 93	97
					<i>G</i>	2 22	1 00	2 22	98 6

School II is outstandingly superior at all points. School I with 39% Italians is second, and School III with 63% Italians is lowest. This does agree closely with the findings of other investigators. We shall make no attempt to explain this phenomenon. Of course, under no circumstances can we make any generalizations concerning *Italian* intelligence. It may well be that the type of employment we have to offer selects out the lowest type of Italian for immigration. On the other hand, it may be that the Italians we have are not dull but that the tests put them at a special disadvantage, linguistically or otherwise.

Returning, however, to our rural-urban comparisons. It would be unfair to draw any conclusions as to rural superiority (on Test B) from the results of tests on Schools I and III. For the reasons outlined above, however, comparison with School II is entirely in order. Table XXVII and Graph VII show the differences between rural District A and urban School II. Whereas these differences are not as marked as in Graph V (Chapter V), still they are consistently and fairly reliably in favor of the rural district. In the former case, the $\frac{\text{Difference}}{\text{Sigma}}$ " is, at its lowest, 5.66, at its highest, 9.33 (where conventional reliability is 3). In the latter case, the average number of chances (in 100) of a true difference is 99.2.

TABLE XXVII.
RURAL AND URBAN SCORES ON INFORMATION B

Reliability of Difference Between Rural District A and Urban School II.*				
Grade and Sex	Actual Difference (Favor of Rural)	σ Difference	Difference in σ units	Chances that true diff. is above 0
4 B	1 52	.79	1 92	97 in 100
G	1 80	.78	2 31	98 9 in 100
5 B	2 24	.77	2 91	99 8 in 100
G	2.64	.87	3 03	100 in 100
6 B	1 44	.63	2.29	98.9 in 100
G	1.80	.71	2 54	99 4 in 100
7 B	2 30	.85	2 71	99.7 in 100
G	3.42	.83	4 12	100 in 100

* Also see Table XXI, Chapter V.

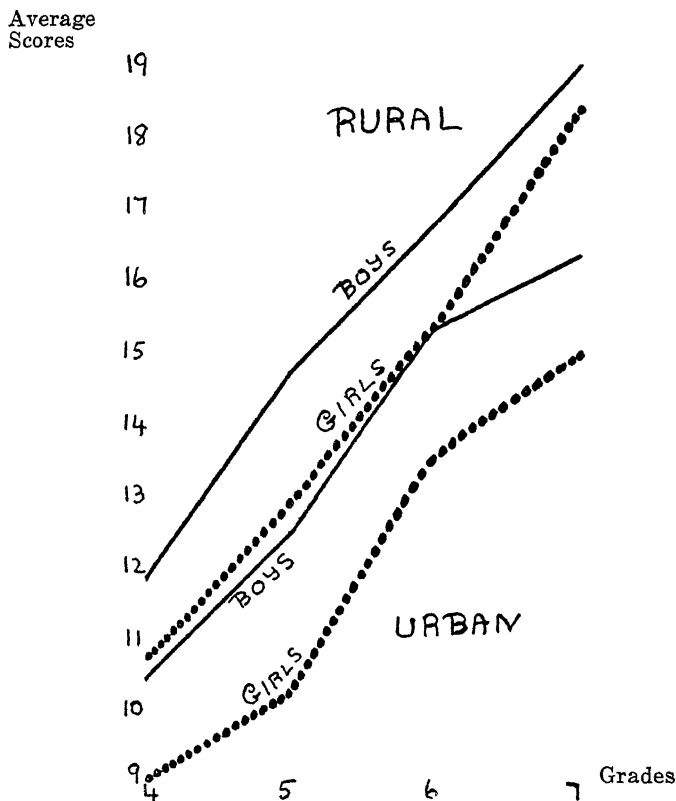


FIG. 7. Information B. Urban (School II) and Rural (District A) Scores

Our conclusions as to urban rural differences on Test B still hold, therefore, when the two groups are equalized as to nationality components.

We have brought to light the striking differences between Schools I, II and III in the same city. Any one of these schools would have provided numbers as large as is used in most norm establishing, but the results in no two of them would have been alike. This is an extremely important point too often lightly passed over by standardizers.

The literature we have summarized has brought out the fact that the different races and nations differ decidedly in their informational and language equipment. It would seem superfluous to mention this except for the fact that conclusions are constantly being drawn without taking this sufficiently into consideration. In extreme cases, as, for exam-

ple, testing primitive man, they seem rather obvious, but when we are testing children who have been in our own schools and who live in our own country, we are too apt to class them all as living in the same environment, whether they dwell on farms or in foreign quarters, or in the shadow of the campus. Porteus and Babcock voice this criticism rather emphatically: "The Army Alpha test, if it does nothing more, proves conclusively how little the psychologists know of the mental range of the man in the street, the laborer on the farm, the mechanic in the shop, in short, of the average man whose intelligence he seeks to measure." . . . "Even so in directions, *e.g.*, Test VII, 'What you are to do in each line is to see what the relation is between the first two words and underline the word in heavy type that is related in the same way to the third word.' For the man whose only idea of relations is of his uncles and aunts, these directions would seem very involved and quite incomprehensible." More so in the vocabulary test: 'allure-attract; deride-ridicule; haggard-gaunt; orifice-aperture' are pairs of words which come quite early in this test and are surely outside the ordinary man's range of understanding. Quite evidently this is no test of intelligence except for those who have had considerable educational opportunities." . . . "How unfortunate that the farmers, the mechanics, the electricians, could not devise a test for the psychologists. If they did, one wonders what the mental age of the latter would appear to be."⁷²

SUMMARY

1. A brief summary of studies on racial and national differences in mentality seemed to point to the conclusion that we are growing more conservative in our interpretations. There is, in many quarters, at least some appreciation of the fact that no dogmatic conclusions can be drawn as to the comparative intelligence of various groups, while our tools and methods are still so imperfect.

2. Analysis of our urban and rural groups, according to the parentage of the subjects, reveals the fact that the rural population is very largely native American while the urban school population varied from 10% to 73% with both parents foreign-born.

⁷² 66, pp. 197-9.

3. School II, with the smallest number of foreign-born, is outstandingly superior on our tests.

4. The children of School II give us a fair basis of comparison with our rural subjects. But the difference on Information B, already elaborated in Chapter V, still holds when the national and racial composition of the two groups is equated.

5. The reliable differences found between the three urban schools (any of which contains a large enough number of cases to satisfy most norm makers) shows us how careful we must be in the selection of our subjects, and the caution we must observe in comparing groups. Mere numbers do not seem to iron out differences.

CHAPTER VII

SUMMARY AND CONCLUSIONS

Our aim, as stated in the introduction, was twofold: 1, to examine the importance of the differentiation of norms according to sex, racial composition, etc., and 2, to ascertain whether mental differences between groups were a function of their innate intelligence or our tests as tools.

In the six preceding chapters we have examined our data from these viewpoints and have arrived at certain conclusions summarized at the end of each chapter. On the whole, five conclusions are outstanding:

1. The adequate scaling and standardization of tests is the first essential of their use as reliable tools.
2. Large numbers, differentiation of the groups as to age, sex, education and natio-racial composition are absolutely necessary. Heterogeneous, hodge-podge groups are useless, if the results on one group are to be compared with those of another.
3. We have found that our rural group fell below the urban group on Information A, scaled on urban children. It is also true, however, that our urban children fell below the rural group on Information B, scaled on rural children. The groups, therefore, are *different* and inferiority is dependent on which group the existing tool favors.
4. Great natio-racial differences were found between our three urban schools. The test scores showed a distinct correlation with these differences. Extreme caution must then be used in the selection of groups for scaling and standardizing tests. We have observed that, in practically all comparisons, tests scaled on one group have been utilized, usually to the disparagement of the second group. Informational and language factors have been shown to be so important that such a procedure is obviously unfair and unscientific.
5. An analysis, from our own point of view, of the literature on national and racial differences leads us to the hypothesis that the groups are *different* rather than that one is *superior* or *inferior*.

It may be said, of course, and with some justice, that we have as much bias towards our interpretation of the facts

as the other investigators whom we have criticized. The reader, then, may take a middle course and steer between the two extremes.

For practical purposes, our findings, even if taken seriously, do not radically change the testing situation. It still is true, for example, that in order to do well in *our* schools a certain grade of performance on *our* tests is necessary. So it is perfectly justifiable to put any individual from any group through our tests and on the basis of the results classify him in a certain way in regard to a specific school, job, etc.

In the same way, we may put one of our own group through an elementary geography test and if he fails we may say with justice that he cannot be admitted to an advanced class. In neither case, however, can we draw any valid conclusions as to innate intelligence.

We have not meant in any way to disparage the value of mental tests. Rather, let it be said that we so firmly believe in their use that we are loath to see their scientific footing jeopardized by the way in which they have been treated by many investigators. At some time in the future we may so perfect our weapons of attack that we may make valid comparisons between different groups. At present, let us frankly admit that we have no adequate method of estimating in any final way the mental differences between the various races and nations. All studies must, then, be very cautiously received and evaluated according to the methods and tools used.

APPENDIX

PRELIMINARY SCALE A

1. What state do you live in?
2. What people were in America when the white men came?
3. What colors are in the rainbow?
4. Why does the beating of your heart keep you alive?
5. What is a submarine boat?
6. What causes an eclipse of the sun?
7. What is the value of the smallest silver coin we use?
8. Who is president of the United States?
9. What is the shortest month in the year?
10. Of what is rubber made?
11. How is it that newspapers can be sold for much less than the cost of printing them?
12. Of what are shoes made?
13. What is the largest city in the United States?
14. Name four different trees.
15. In what country is Vienna?
16. What is the usual economic result of the over-production of any commodity?
17. What is our national song?
18. Why should we kill flies?
19. In what month of the year do the days begin to grow shorter?
20. How can banks afford to pay interest on the money you deposit?
21. Why is it dark at night?
22. What form of government have we in the United States?
23. Name at least three bodily processes that go on when a person reading music plays the piano?
24. What is the shape of the earth?
25. Name three precious stones?
26. Why is the moon light at night?
27. How much does it cost to mail a letter to any city in the United States?
28. Of what is butter made?
29. What artificial water way connects the Atlantic with the Pacific?
30. Who was the first president of the United States?
31. Where does the sun rise?
32. Name a country in Europe which is a republic.
33. How do you know a policeman when you see him?
34. How many states are there in the United States?
35. Who is the governor of your state?
36. What is the function of respiration?
37. What may we expect when we see heavy black clouds?
38. What do the stars in the American flag represent?
39. What is a referendum in government?
40. Name five vegetables.
41. What is steam?
42. What is the economic value of Alaska to the United States?
43. What is the capital of the United States?
44. Why do we celebrate the Fourth of July?
45. What are the functions of the three branches of our government? (In three words.)
46. How many months are there in a year?
47. What three things do most plants need in order to live?
48. For how many years is the president of the United States elected?
49. What is the largest river in the United States?
50. In your city what is the youngest age at which a child can leave school?
51. Name the greatest English writer of plays.
52. Of what use are insects to flowers?
53. To what public building can you go for books?
54. Where does Congress meet?

55. What is a civil war?
56. How many cents are there in a quarter?
57. Why don't we see the stars in the daytime?
58. Why did the Pilgrims come to this country?
59. How many sides has a triangle?
60. How can you tell when water is boiling?
61. Who was President of the United States during the World War?
62. What holiday comes in December?
63. What part of the night or day is 12.30 A. M.?
64. What is the freezing point of water?
65. How many hours are there in a day?
66. How old must you be before you can vote?
67. Of what is paper made?
68. Name the Great Lakes.
69. What is vaccination for?
70. About how often do we have a full moon?
71. Name four general reasons that prevent a would-be immigrant from entering the United States.
72. What are the colors in the American flag?
73. What holiday do we now celebrate that was first celebrated by the Pilgrims?
74. Name five insects.
75. Name three wars in the United States fought with other countries.
76. How many weeks are there in a year?
77. Name two stones used for building purposes.
78. How many pints are there in a quart?
79. Name five cities of the United States that have a population of over half a million.
80. Why are there no shadows on a heavily clouded day?

KEY TO SCALED INFORMATION A

1. *Red*, white and *blue*.
2. Leather, satin, wood, etc. (Any sensible answer is accepted.)
3. Twelve or twenty-four.
4. Christmas.
5. President at time of test.
6. Indians, red men.
7. State child lives in.
8. East.
9. Any *five* vegetables. (No partial credits allowed.)
10. Mississippi or Missouri-Mississippi.
11. Religious toleration, freedom from persecution, right to worship God in own way. (Any phrasing of above idea is accepted.)
12. Four years.
13. Sunshine or warmth, water or rain, soil or ground, air. Any three of the preceding (or their chemical components). Only one member of a pair accepted.
14. Circulates blood, sends blood through veins, pumps blood. (Any phrasing of this idea accepted.)
15. They loan it at higher interest, they buy mortgages. (Any answer that includes idea of *using* money accepted.)
16. Any *five* insects.
17. War between two parts of same country, same countrymen fighting, brother against brother, etc.

18. At least *three* of *starred* cities below and *two* of the others: (Four starred and one other, or five starred.)¹

Akron	Columbus	Minneapolis	Providence
Atlanta	*Detroit	Milwaukee	Rochester
*Baltimore	Denver	Newark	St. Paul
*Boston	Indianapolis	*New York	*St. Louis
*Buffalo	Jersey City	Oakland	*San Francisco
Brooklyn	Kansas City	*Philadelphia	Seattle
*Chicago	*Los Angeles	*Pittsburgh	Toledo
Cincinnati	Louisville	Portland, Ore.	Washington
*Cleveland	*Manhattan		

19. Any answer that mentions advertising. (Large production not accepted.)
 20. 32 or 32 Fahrenheit or 0 Centigrade. (Zero not accepted.)
 21. Higher prices. (Waste, panic, etc., accepted.)
 22. Executive or administrative, judicial, legislative. (Any phrasing of these ideas accepted.)
 23. Quota exhausted, defective mentality, contagious disease, lack of funds, moral turpitude, criminality, beliefs contrary to United States Constitution, bigamy, member of excluded race. Any four expressions of above accepted. (Any doubtful answer should be looked up in latest immigration rulings.)
 24. To supply blood with oxygen, to carry off waste material, etc.
 25. Any answer that includes the idea of a legislative measure's being referred to people.

¹ Although the question calls for five cities with a population of over half a million, it was marked correct if three such cities (starred above) were given, together with three other large cities from above list.

PRELIMINARY SCALE B

- What is our national song?
- Name the young of the sheep, cow, horse.
- Why don't we see the stars in the daytime?
- Name *three* states in the U. S. where cotton is raised.
- Name *four* different trees.
- What is the correct temperature for a living room?
- In what month of the year do the days begin to grow shorter?
- How many cents are there in a quarter?
- What makes cobwebs?
- Where does Congress meet?
- Why is it necessary to limit the hunting season?
- Of what is butter made?
- In what part of the day are the shadows longest?
- Why is the moon light at night?
- Draw a square and an oblong.
- What is the capital of the U. S.?
- Why do unsupported objects fall to the ground?
- What is the largest city in the U. S.?
- How can you tell poison ivy by looking at it?
- About how often do we have a full moon?
- How many pecks are there in a bushel?
- What is the shortest month in the year?
- Name *three* different plants from which sugar is made.
- Of what use are insects to flowers?
- What is the youngest age at which a child can leave school?
- Name a famous American inventor and tell what he invented.
- Name *two* stones used for building purposes.
- What are the four seasons?
- How many states are there in the U. S.?
- Name *five* wild flowers.
- Why does seasoned wood burn more easily than green wood?

32. What may we expect when we see heavy black clouds?
33. Tell one way of finding out the age of a tree.
34. Why does frost form on the *inside* of the window pane?
35. How many sides has a triangle?
36. From what animal do we get mutton?
37. What kind of cloth is made from flax?
38. From what does maple sugar come?
39. Who was the first President of the U. S.?
40. Name *two* birds that stay North in the winter.
41. How much does it cost to mail a letter to any city in the U. S.?
42. What do we mix with ice to help us freeze ice-cream more quickly?
43. How do sponges grow?
44. How can you keep milk from souring?
45. What do the stars in the American flag represent?
46. Name the continents in order of size.
47. Give one reason for the rotation of crops.
48. What holiday do we celebrate that was first celebrated by the Pilgrims?
49. Name *five* crops.
50. What causes an eclipse of the sun?
51. What kind of dairy cow gives the richest milk?
52. What is a submarine boat?
53. Why should we kill flies?
54. How many pints are there in a quart?
55. Why are crops hoed?
56. What is steam?
57. At what time of year do many leaves turn red?
58. Name *three* products made from wheat.
59. What form of government have we in the U. S.?
60. Name a vegetable that grows above ground.
61. What is a domestic animal?
62. Why do we celebrate the 4th of July?
63. Of what is paper made?
64. Who is the Governor of your State?
65. What artificial waterway connects the Atlantic with the Pacific?
66. How many months are there in a year?
67. Why is it dark at night?
68. How can you locate the Pole star?
69. How old must you be before you can vote?
70. Name a country in Europe which is a Republic.
71. What is the highest court in the U. S. called?
72. Of what is rubber made?
73. Who was the President of the U. S. during the World War?
74. Name *two* animals that hibernate in winter.
75. What is the economic value of Alaska to the U. S.?
76. What tree doesn't shed its leaves in the Fall?
77. Name *two* differences between the barks of birch and oak trees.
78. Name *three* uses of forests.
79. Name *five* fruits.
80. How many weeks are there in a year?

PRELIMINARY TEST—SCALED INFORMATION B

DIRECTIONS FOR ADMINISTRATION

1. There are three sheets of this preliminary test. Difficult and easy questions are scattered throughout. We are attempting to test the questions rather than the children. So it is very important that each question be given due consideration and that enough time be given for all the questions. We hope that you can give these tests (page by page) directly you receive them; *i.e.*, the first morning, questions 1-25; the second morning, questions 26-53; the third morning, questions 54-80. Fifteen minutes for each page will probably be sufficient.

2. Say to the children: "Some people are trying to find out what sort of questions boys and girls of your age can answer. They have asked us to help them. I'm going to give each of you a set of questions. Don't be discouraged because you can't answer all the questions, but be sure to do as many as you can. Try every question on the whole page. You are not going to be marked on your answers. In fact, you don't have to put your name on the paper at all. Do not start on the test until I tell you to do so."

3. Distribute the tests, one to each child. Have the children write on one of the top lines (in the box in the upper right hand corner) *Boy* or *Girl*. Be sure that each child gives his age in years and months, and his school grade.

4. When this has been completed, give the signal to commence the test. Do not give any help whatsoever. If you observe any child cheating, destroy his paper.

SCALED INFORMATION B

DIRECTIONS FOR ADMINISTRATION

1. Please give this test on the morning of May 1 (or as near then as possible).

2. Say to the children: "Some people are trying to find out what sort of questions boys and girls of your age can answer. They have asked me to help them. I am going to give each of you a set of questions. I want you to look at each one carefully and to write the proper answer in the space beside each question. Don't be discouraged because you can't answer all the questions, but be sure to do as many as you can. You are not going to be marked on your answers. In fact, you don't have to put your name on the paper at all. Do not start on the test until I tell you to do so."

3. Distribute the tests, one to each child. Have the children write in the box in the upper right-hand corner, *Boy* or *Girl*. Be sure that each child gives his age in years and months, and his school grade.

4. When this has been completed, give the signal to commence the test. Allow *15 minutes only*. Do not give any help whatever. If you observe any child cheating, destroy his paper.

5. After fifteen minutes give the signal to stop. Have every child turn his paper over. Tell each pupil to write the name of the country where he was born; underneath that, the name of the country where his father was born; and underneath that, the name of the country where his mother was born.

KEY TO SCALED INFORMATION B

1. Cream, milk, whey, cream and salt.
2. Two cents.
3. Spring, summer, autumn (fall), winter.
4. "Star Spangled Banner," "America," "My Country, 'tis of thee," "Oh, say, can you see?"
5. Squash, lettuce, tomato, cabbage, string beans, peas, cucumber, etc.
6. Lamb, calf, colt (pony, foal, filly).
7. Salt.
8. Any *four* different trees.
9. Sheep.
10. Dry, dryer, green wood is wet, crisp, contains less water.
11. Any *five* crops.
12. Pine, hemlock, fir, spruce, liveoak, cedar, palm, Christmas tree, evergreen, balsam.
13. Annual rings, rings, lines or marks in stump, cambium layers, layers in the trunk, size, height, tallness, extent of the roots, roughness of the bark, etc.

14. Light from the sun, obscured by the sun, daylight too strong. (Any answer mentioning brightness.)
15. Every four weeks, or month, or 28 or 31 days. Twelve times a year, etc.
16. Any *five* flowers.
17. Fisheries, minerals, furs or valuable skins, gold, coal, salmon, seals. Any money value over \$5000 a year.
18. Condensation of moisture, warmer inside, steam or vapor inside, heat in the room, heat and cold come together. Relative humidity of air inside and out.
19. Lumber, wood, fuel, draw water, prevent floods, keep soil in good condition, prevent landslides, shade, scenery and beauty, recreation, camping and health resorts, animal refuge or protection, hunting, trapping, home for birds, tar, petroleum, cork, dyes, medicines or chemicals from trees, maple sugar or sap, windbreaker, regulate temperature, etc. (Any *three*.)
20. Wilson.
21. Replacement of nitrogen, fertilization, different crops use up different substances in the soil, keep the soil from becoming exhausted, better crops, etc.
22. Sugar cane, beet, maple, sorghum, palm, corn, grapes. (Any *three*.)
23. June.
24. Reflection of the sun, light from the sun, sun shines on it, etc.
25. Big dipper, dipper, Great Bear, pointer stars, latitude, compass.

BIBLIOGRAPHY

Information

1. Eastman, E. An Information Test Applied to Juvenile Delinquents. *J. Appl. Psych.*, 1926, 10, 202-215.
2. Pressey, S. L., and Shively, I. M. A Practical Information Test for Use with Delinquents and Illiterate Adults. *J. Appl. Psych.*, 1919, 3, 374-380.
3. Terman, L. *Genetic Studies of Genius*, Vol. 1. Stanford Univ. Press, 1925.
4. Thorndike, E. L. *The Measurement of Intelligence*. N. Y. Bureau of Pubs. T. C., New York, 1926.
5. Weeks, A. L. A Vocabulary Information Test. *Arch. of Psych.*, 1928, 97.
6. Whipple, G. M. *A Manual of Mental and Physical Tests*, Vol. II. 1915, Balto., Warwick & York.

Age and Sex Differences

7. Book, W. F., and Meadows, J. L. Sex Differences in 5925 High School Seniors in Ten Psychological Tests. *J. Appl. Psych.*, 1928, 12, 56-81.
8. Bronner, A.; Healy, W.; Lowe, G.; Shimberg, M. *Manual of Individual Mental Tests and Testing*. Boston, Little, Brown, 1927.
9. Brooks, F. D. Rate of Mental Growth, Ages 9-15. *J. Educ. Psych.*, 1921, 12, 502-510.
10. Goodenough, F. L. The Consistency of Sex Differences in Mental Traits at Various Ages. *Psych. Rev.*, 1927, 34, 440-463.
11. Hall, G. S. The Contents of Children's Minds on Entering School. *Ped. Sem.*, 1891, 1, 138-173.
12. Hollingworth, L. S. Comparison of the Sexes in Mental Traits. *Psych. Bull.*, 1918, 15, 427-432.
13. Hollingworth, L. S. Sex Differences in Mental Tests. *Psych. Bull.*, 1916, 13, 377-384.
14. King, I., and McRory, J. Freshman Tests at the State University of Iowa. *J. Educ. Psych.*, 1918, 9, 32-46.
15. Lincoln, E. A. Sex Differences in American School Children. 1927, Balto., Warwick & York.

16. Pressey, L. W. Sex Differences Shown by 2544 School Children. *J. Appl. Psych.*, 1918, 2, 323-340.
17. Vorstellungskreis der Berliner Kinder beim Eintritt in die Schule. *Berlin Städtisches Jahrbuch*, 1890, 59-77.
18. Whipple, G. M. Sex Differences in Intelligence Test Scores in the Elementary School. *J. Educ. Res.*, 1927, 15, 111-117.
19. Winsor, A. L. The Relative Variability of Boys and Girls. *J. Educ. Psych.*, 1927, 18, 327-336.
20. Woolley, H. T. The Psychology of Sex. *Psych. Bull.*, 1914, 11, 363-379.

Rural

21. Abel, J. F. A Study of 260 School Consolidations. *Bull.*, 1924, 32. Washington, Government Printing Office.
22. Baldwin, B. T., and Fillmore, E. L. The Mind of the Rural Child. *Psych. Bull.*, 1928, 25, 185-186.
23. Bickersteth, M. E. Application of Mental Tests to Children of Various Ages. *Brit. J. Psych.*, 1917, 9, 23-73.
24. Book, W. F. The Intelligence of High School Seniors. New York, Macmillan Co., 1922.
25. Brunner, E. de S. Village Communities. New York, Doran, 1927.
26. Cook, K. M. Distribution of Consolidated and One-Teacher Rural Schools. *J. Rural Educ.*, 1925, 4, 337-347.
27. Chapman, J. C., and Eby, H. L. A Comparative Study, by Educational Measurements, of One-Room Rural School Children and City School Children. *J. Educ. Res.*, 1920, 2, 636-646.
28. Duff, J. F., and Thomson, G. H. The Social and Geographical Distribution of Intelligence in Northumberland. *Brit. J. Psych.*, 1923, 14, 192-198.
29. Foote, J. M. Comparative Study of Instruction in Consolidated and One-Teacher Schools. *J. Rural Educ.*, 1923, 2, 337-351.
30. Frost, N. A Comparative Study of Achievement in Country and Town Schools. *T. C. Contrib. to Educ.*, 11, New York, 1921.
31. Gray, P. L., and Marsden, R. E. Intelligence Tests in Rural Schools. *J. Exp. Ped.*, 1922, 6, 224-231.
32. Hinds, J. H. Comparison of Brightness of Country and City High School Children. *J. Educ. Res.*, 1922, 5, 120-124.
33. Hollingworth, L. Gifted Children. New York, Macmillan, 1926.
34. Irion, T., and Fisher, F. C. Testing the Intelligence of Rural School Children. *Amer. Schoolmaster*, 1921, 14, 221-223.
35. Lehman, H. C., and Witty, P. A. The Psychology of Play Activities. New York, A. S. Barnes, 1927.
36. Myers, C. E. Measuring Educational Efficiency. *Res. Bull. Penn. State Educ. Assoc.*, 1928, 3.
37. Pintner, R. A Mental Survey of the School Population of a Village. *Sch. and Soc.*, 1917, 5, 597-600.
38. Pressey, L. W. The Influence of Inadequate Schooling and Poor Environment Upon Results with Tests of Intelligence. *J. Appl. Psych.*, 1920, 4, 91-96.
39. Pressey, S. L., and Thomas, J. B. A Study of Country Children in a Good and a Poor Farming District by Means of a Group Scale of Intelligence. *J. Appl. Psych.*, 1919, 3, 283-286.
40. Pyle, W. H., and Collings, P. E. Mental and Physical Development of Rural Children. *Sch. and Soc.*, 1918, 8, 534-539.
41. Reisner, G. A Study of Certain Educational Tests of Mental Ability to Determine Whether or Not There Is Any Relation Between Type of Question and Score Made by Urban and Rural Pupils. M. A. thesis (unpub.), Penn State College, 1927.
42. Thomson, G. H. The Northumberland Mental Tests. *Brit. J. Psych.*, 1921, 12, 201-222.
43. Smith, W. C. The Rural Mind. *Amer. J. Soc.*, 1927, 32, 771-786.

Rural Surveys

44. Public Education in *North Carolina*. New York General Educ. Board, 1921.
45. *Indiana*—Rural Education. Survey Committee, 1926.
46. Public Education in *Kentucky*. New York, General Educ. Board, 1922.
47. O'Shea, M. V. A State Educational System at Work (*Mississippi*). 1927.
48. Works, G. A. Rural School Survey of *New York State*. N. Y., Ithaca, 1922.
49. *Texas Educational Survey*, 1925.
50. *Virginia*, Public Schools, Pt. II. New York, World Book Co., 1921.

Statistical References

51. Garrett, H. E. Statistics in Psychology and Education. New York, Longmans, Green, 1926.
52. McCall, W. A. How to Experiment in Education. New York, Macmillan 1923.
53. McCall, W. A. How to Measure in Education. New York, Macmillan, 1929.
54. Rugg, H. O. Statistical Methods Applied to Education. Boston, Houghton, Mifflin, 1917.

Racial and National

55. Armstrong, C. P. A Study of the Intelligence of Rural and Urban Children. (Unpublished to date.)
56. Davis, R. A. Some Relations Between Amount of School Training and Intelligence Among Negroes. *J. Educ. Psych.*, 1928, 19, 127-130.
57. Garth, T. R. The Intelligence of Full Blooded Indians. *J. Appl. Psych.*, 1925, 9, 382-389.
58. Helmer, V. The American Indian and Mental Tests. M. A. Thesis (unpub.), Univ. of Kansas, 1926.
59. Herskovits, M. J. The Negro and the Intelligence Tests. *Ped. Sem.*, 1926, 33, 30-42.
60. Hirsch, N. D. A Study of Natio-Racial Mental Differences. *Gen. Psych. Monog.*, 1926, 1, 231-406.
61. Jones, V. A. A Study of the Non-Verbal Nature and Validity of the Myers Mental Measure. *J. Educ. Res.*, 1927, 16, 203-209.
62. Kirkpatrick, C. Intelligence and Immigration. *Ment. Meas. Monog.*, 1926, 1.
63. Klineberg, O. An Experimental Study of Speed and Other Factors in "Racial" Differences. *Arch. Psych.*, 1928, 93.
64. Koch, H. L., and Simmons, R. A Study of the Test Performance of American, Mexican and Negro Children. *Psych. Monog.*, 1926, 5.
65. Peterson, J. Comparative Abilities of White and Negro Children. *Comp. Psych. Mon.* 1923, 1.
66. Porteus, S. D., and Babcock, M. E. Temperament and Race. Boston, Badger Co., 1926.
67. Sunne, D. Comparison of White and Negro Children in Verbal and Non-Verbal Tests. *Sch. and Soc.*, 1924, 19, 469.
68. Sweeney, A. Mental Tests for Immigrants. *No. A. Rev.*, 1922, 215, 600-612.
69. Wang, S. L. Demonstration of Language Difficulty Involved in Comparing Racial Groups by Means of Verbal Intelligence Tests. *J. Appl. Psych.*, 1926, 10, 102-106.
70. Wells, G. R. The Application of the Binet-Simon Tests to White and Colored School Children. *Psych. Monog.*, 1923, 32, 52-58.
71. Whitney, F. L. Intelligence Levels and School Achievement of the White and Colored Races in the United States. *Ped. Sem.*, 1923, 30, 69-86.
72. Yoder, Dale. The Present Status of the Question of Racial Differences. *J. Educ. Psych.*, 1928, 19, 463-470.

VITA

Myra Esther Shimberg. Born, Troy, N. Y., 1901. Was educated in France and England until 1915. B.A., Wellesley, 1922. Phi Beta Kappa, 1922. M.A., Wellesley, 1924. Assistant in Psychology at Wellesley for one year. On psychological staff of Judge Baker Foundation, Boston, for three years as clinical and research worker. Joint author of "A Manual of Individual Mental Tests and Testing." Published articles in The American Journal of Psychology and The Journal of Applied Psychology. Alice Freeman Palmer Fellow, 1927-8. Studied at Columbia, 1927-9. Sigma Xi, 1929.

BLOOD PRESSURE CHANGES IN DECEPTION

BY
MATTHEW N. CHAPPELL

Submitted in partial fulfilment of the requirements for the degree
of Doctor of Philosophy, in the Faculty of Philosophy,
Columbia University.

NEW YORK
April, 1929

ACKNOWLEDGMENTS

The author wishes to take this opportunity to thank Dr. Joseph L. Holmes under whose direction this work was undertaken; Professor R. S. Woodworth and Professor A. T. Poffenberger for valuable criticisms and suggestions; the Taylor Instrument Company for the use of the Recording Sphygmomanometer; and the students of the department of psychology for their loyal support.

TABLE OF CONTENTS

I	Introduction.	
	A critical survey of the work of Marston and of Landis	5
II	Apparatus	7
III	The Circulatory System.....	8
IV	A Comparison of the Methods of Obtaining Lateral Blood Pressure	12
V	Detection of Deception by Increases in Lateral Blood Pressure	14
VI	Control Experiments.....	20
	A. Rises during an "intelligence test.".....	20
	B. Rises due to mental work.....	23
	C. Rises due to unemotional deception.....	25
VII	Summary and Conclusions.....	27

BLOOD PRESSURE CHANGES IN DECEPTION

The detection of deception by means of physiological changes was first undertaken by Benussi¹ in 1914. He found that changes in respiration usually indicated deception. No one followed Benussi's lead until Burt² in 1921 repeated it and somewhat improved the technique. In the meantime Marston,³ then at Harvard, conceived the idea of using blood pressure as a criterion of deception. His technique, which we shall discuss later in this paper, was essentially that employed by Benussi in the earlier work. Burt, in 1921² repeated the work of Marston and found the blood-pressure method highly reliable. Landis and Gullette,⁴ in 1926 could find no evidence that the pressure changes were of value in detecting deception. Careful reading reveals that the differences between the latter work and that of Marston are not as great as they seem at first glance.

In his investigations Marston used the Tycos Indicating Sphygmomanometer and the auscultatory method of taking pressure readings. His situation was one in which the subject might assume himself to be guilty or innocent of an alleged crime. In working up the data he treated each subject individually. The rise in pressure which he uses is the highest reading above the initial pressure. He concludes from his results:

- (1) That the consciousness of an attitude of deception is the key to the changes in pressure.
- (2) That a rise of 12mm. is significant of deception.
- (3) That deception can be detected by the pressure change with accuracy of 90-100%.
- (4) That "a uniform and significant systolic pressure curve was established by the results as symptomatic of the

¹ Benussi, V. (*Archiv. fur die gesamte Psychol.*) 1914, 244-273 (47).

² Burt, H. E. (*Jour. Exp. Psych.*) 1921, IV 1-23 (27), (49), (58).

³ Marston, W. (*Jour. Exp. Psych.*) 1917.

⁴ Landis & Gullette (*Jour. Comp. Psych.*) 1926.

deceptive consciousness. A rather surprising secondary result was the appearance of an almost equally definite truth curve."

This last conclusion is open to question in the light of the two individual curves given in his report, those of subjects C and G. There is certainly no uniformity apparent in these either in height of the rise or the shape of the curve. The only similarity is that they both do show a rise.

Landis and Gullette also used the Tycos instrument and the auscultatory method. Their subjects were accused of a hypothetical crime and tested twice, once when they assumed themselves guilty and again when they assumed themselves to be innocent. They treated their data not individually but found group changes. All the truth readings for all the subjects were added together and the group average for truth was determined. The data for deception were treated in the same manner and these two group averages were subtracted to find the difference in pressure between truth and deception. This difference was found to be 3mm. The accuracy of the method in differentiating truth from deception was found to be 82%. They conclude from their data "that it was impossible to set up difference in pressure as a criterion of differentiation between truth and falsehood."

The results of Landis and Gullette cannot be compared with those of Marston for the experiments have little in common. The former made their situation as artificial as possible, eliminating emotion instead of injecting it into the situation as Marston did. The treatment of their results was very different from that of Marston. By averaging their group data they have effectively blanketed all individual changes. Further, the difference of the group averages can be readily controlled by the length of the individual test since the excitement in so artificial a situation will be of short duration. The blanketing effect on large changes will be shown later in this report. Their data as taken are not reported. It is, therefore, not possible to determine accurately what the effect of individual treatment would be, but some indication may be gained by inspection of the individual curves presented. These are all grouped in two charts. They show the subject's variability to be somewhat greater when lying than while

telling the truth. In the truth series only three of the twenty-two subjects show rises of 8mm. or more, whereas in the deception series eight subjects show such a rise. Finally, their conclusion that it is impossible to set up pressure differences as diagnostic of deception is not justified by their own results, which show the chances of a reliable difference to be eighty-two in one hundred. When all things are considered it seems quite possible that if Landis and Gullette had used individual treatment of their data their results might have given strong support to some of Marston's contentions.

Following the work of Marston, Larson at the Police School in Berkeley, California devised a method of detecting deception in which he obtained continuous records of circulatory changes in the arm. Just what phenomenon of circulation is recorded by this instrument is not clear. It is a high pressure plethysmograph and records volume changes in the arm of a type somewhat different from the volume changes recorded by the low pressure plethysmograph. That it does not record lateral pressure is shown by data obtained by the present experimenter which are as yet unpublished. For this reason we shall not consider the work of Larson in this paper.

APPARATUS

Two types of instruments were used in this investigation—the Tycos indicating and the Tycos recording sphygmomanometers. A conical stethoscope was used with the indicating instrument to hear the sound of flow through the brachial artery. The Tycos indicating sphygmomanometer is the instrument used by Marston and Landis. It has a single chamber cuff which may be adjusted to the arm or leg. When this is inflated the pressure is indicated on a dial by means of a pointer attached to an aneroid diaphragm. This instrument only indicates and does not record any pressure. Its use involves an observer and the accuracy of the readings obtained varies somewhat with variations of his acuity of hearing.

The second instrument is described in a booklet published by the Taylor Instrument Company. It is purely mechanical, easily operated and cannot be greatly influenced by the observer. With this a record is made in ink on a rotating paper disc similar to those used in taking continuous records of air

and steam pressure in industrial plants except that the pressure is indicated circumferentially rather than radially. The disc is driven by the movement of an aneroid diaphragm which covers the pressure chamber. This chamber is also connected directly to the cuff. In this fashion the chart shows at all times the cuff pressure. This cuff has two chambers so connected that as the air slowly escapes after inflation, the upper chamber is always at a slightly lower pressure than the one below. The pressure in the upper chamber is the same as that behind the aneroid diaphragm which controls the position of the chart. The lower chamber is connected to another aneroid diaphragm which drives the pen. Due to the difference of pressure in the two chambers of the cuff the blood passes the upper chamber at the point of the lateral pressure, which just counterbalances the cuff pressure, and surges against the lower chamber. This causes a slight increase in the pressure in this chamber. The surge is transmitted to the pen through the aneroid diaphragm causing it to move radially. The first radial line appearing on the chart is the point of lateral pressure. As the air leaks out of the cuff the amplitude of the pen increases for a period and then decreases. Usually there is a quick decrease and sometimes a point appears at which two of the radial lines are of about equal amplitude. This has been taken empirically by the Taylor Instrument Company to be the point of diastolic pressure. Unfortunately the criterion of diastolic pressure does not appear in all subjects. The instrument makes a record only of the intermittent phenomena and does not take a continuous record of the changes. Several records may be made on one chart and the curve of the changes plotted directly on it.

THE CIRCULATORY SYSTEM

To understand the variety of changes which may occur in blood pressure it is necessary to have some knowledge of the vascular system. Let us, therefore, briefly point out the major factors governing the changes in the normal individual.

The heart, with the arteries and veins forms a closed system in which circulates a comparatively constant amount of

blood at an average pressure of about 110 to 140mm. of mercury, depending upon the age.

In speaking of blood pressure it is necessary to specify what pressure is meant since there are five different kinds: namely, systolic, diastolic, pulse, mean and lateral. By systolic is meant that pressure which is developed during the systole or the contraction of the heart which drives the blood from the left ventricle in to the general system. By diastolic is meant the pressure in the arterial system at the closure of the semilunar valves at which time occurs the diastole or expansion of the heart. By pulse pressure is meant the difference between the diastolic and the systolic or that portion which drives the blood through the system. By mean is indicated the average of the systolic and diastolic pressures. By lateral is meant the pressure against the arterial walls at any point.

Some confusion has arisen from the failure of investigators to indicate and properly define the pressure with which they have worked. Particularly have systolic and lateral pressures been confused. It is impossible to obtain systolic blood pressure except by the use of the manometer and the stromuhr simultaneously; that is, there are two components in which the energy of the contraction is expended, namely, velocity head and pressure head.⁵ The stromuhr measures the velocity and the manometer the pressure head. Since the use of these instruments involves an incision in the arteries, needless to say they are not used with human subjects. Still many papers may be found purporting to deal with systolic pressure. If we could use systolic pressure the problem would be relatively simple since we would not be dealing with those factors outside of the heart which cause the lateral pressure to vary. That is, we would be concerned only with the decrease and increase of the heart activity and could eliminate the influence of the vaso-motor system. What passes for systolic pressure in the human being is more properly lateral pressure. To call this systolic is misleading for it is not necessarily directly proportional to the pressure of the systole. Since it is with lateral pressure that we are dealing let us consider just what factors influence its change.

⁵ Burton-Opitz,² Textbook of Physiology.

First, it may be influenced by an increase or decrease in the force of the heart beat or, force remaining constant, by the rate of the beat. Second, it may be raised by any major vaso-constriction or lowered by any major vaso-dilation. Third, a change in the quantity of blood circulating may raise or lower the pressure depending upon the direction of the change. The third factor is of no importance to our investigation since over one quarter of the total amount of blood must be lost before any marked change in pressure occurs. So far as we know none of our subjects lost any blood.

Those effects coming directly from the heart action are governed by two sets of nerve fibers which originate in the medulla.⁶ These are called the augmentor and inhibitor fibers. They are efferent in nature. The augmentor fibers leave the spinal cord in the anterior roots of the second and third and probably the fourth thoracic nerves. They reach the sympathetic chain through the white rami communicantes running up through the stellate ganglion and the annulus of Vieussens to the inferior cervical ganglion. From here they pass off to the heart by separate accelerator branches. The inhibitor fibers leave the medulla in the roots of the eleventh, or spinal accessory nerve which joins the roots of the tenth or vagus. They pass down the vagus in two bundles, one through the superior and the other through the inferior cardiac ramus to the plexus cardiacus located on the arch of the aorta where also are found the augmentor fibers.

Another important nerve directly connected with the heart is the depressor. It is sensory in its nature and conducts impulses from the arch of the aorta to the nucleus of the vagus and the cardiac and vasomotor centers. It gives rise to reflexes which produce a reduction in the rate of the heart beat and a dilatation of the blood vessels. It probably functions only in emergencies.

The augmentor fibers function in two ways: first to increase the rate and second, to increase the force of the heart beat. These are independent factors, the one occurring with or without the other. Excitation of the inhibitor fibers causes a decrease in the activity of the heart. The nature of this reduction is not clearly understood. It seems probable,

⁶ Stewart, Manual of Physiology.

however, that the activity of the vagus causes a liberation of some salt, possibly potassium, in the heart itself which raises the synaptic resistance or to some extent anaesthetizes the sympathetic nerve endings, thus decreasing the augmenting impulses.

If we were considering systolic blood pressure these might be the only parts of the nervous system which would interest us. However, since we are dealing with lateral pressure we must consider the excitation of the vasomotor mechanism which in itself is in a state of constant change. The vasomotor center is located in the medulla in the same region as that in which the cardiac center is found.⁷

Just as the muscular walls of the heart are governed by two sets of nerve fibers, a set which keeps down the rate of working and a set which may increase it, the muscular walls of the vessels are under the control of nerves which have the power of diminishing their size (vaso-constrictor) and those which have the power to increase it (vaso-dilator). These vasomotor nerves control chiefly the small arteries. They probably have no direct influence on the capillaries. The fibers seem to arise from any part of the spinal cord, the constrictors, being somewhat more restricted than the dilators. The chief vasomotor nerves are the sciatic, the brachial, the trigeminus, the cervical sympathetic, the greater splanchnic and on the afferent side only the depressor.⁸

In order to get a clearer conception of the integrated action of the nervous mechanism upon the circulation let us take a few hypothetical conditions. Assume the force of the heart to be F and the rate to be R and the vasomotor condition to be M (in the arm of a subject), the lateral pressure to be L and the velocity pressure to be V :

- (1) If now R and M remain constant and F increases it results in an increase of L and V , that is, the lateral pressure which we record is higher.
- (2) If F and R remain constant and M becomes C (constriction) there is a greater obstruction to the flow and a decrease in V . With a decrease in V the energy

⁷ Ranson, *Am. Jour. Physiol.*, 1917.

⁸ Stewart, *Manual of Physiology*.

loss in V is converted into lateral pressure and there is, therefore, an increase in L . Again then, we record a higher lateral pressure and this time with no increase in the force of the heart beat which is systolic pressure.

- (3) If F increases and R remains constant and M becomes D (dilation) just sufficient to keep V constant, then L also will remain constant and we will have a condition in which there is a rise in systolic pressure and no rise in lateral pressure.
- (4) Many other combinations may be demonstrated. Probably the most extreme would be that in which an increase in F with a general vaso-dilation may result in a drop of lateral pressure.

The above is a simplified view of the phenomena, which probably never occurs, that is, one factor does not change while all the rest remain constant but all may change simultaneously and in any direction. For a more complete explanation of the circulation the reader may peruse any good text book of physiology.

THE COMPARISON OF METHODS OF OBTAINING LATERAL BLOOD PRESSURE

Lateral blood-pressure may be taken by three methods, namely, the mechanical method, the auscultatory method and the palpation method.

The mechanical method is that in which the record of the pressure is obtained by purely mechanical means and is uninfluenced by the observer. The auscultatory is that in which a stethoscope is used to observe the sound caused by the flow of blood through some artery, usually the brachial. The palpation method is that in which the flow of blood is observed by placing the finger tips over the radial artery.

Previous investigators have not used the mechanical method. In order to compare work done by the different methods it is necessary to determine the relations existing between them. In this determination the Tycos Recording Sphygmomanometer was used as the standard since it is

mechanically so designed that the record of lateral pressure must be taken when blood flows past the upper cuff of the arm band.

Two series of readings were taken. In the first, simultaneous readings were obtained by the mechanical and the palpation methods, and in the second by the mechanical and auscultatory methods. All the readings in each series were taken on one subject, the interval between readings being about one minute.

The results obtained in the first series are shown in Table I and those in the second in Table II. Table I shows that the average of the thirty readings obtained by the mechanical method is 125mm. Hg. The palpation average is 107.4mm. Hg. The average difference between the two methods is 17.6mm. Hg. This means that the Tycos Recording instrument indicates the lateral pressure 17.6mm. higher than do the finger tips of this observer. This difference will vary among observers, depending probably on the sensitivity of the fingers. The σ_{dis} of the mechanical readings is 5.77. The σ_{ave} is 1.15. For the palpation method the σ_{dis} is 5.08 and the σ_{ave} is .945. The σ of the difference between the two is 1.43. The reliability of the difference is 12.3σ .

It seems probable that with practice in taking the pressure by the pulse method this difference might be somewhat reduced. These readings were taken by an observer relatively unpracticed in all the methods.

Table II shows that for the mechanical method the average is 123.2mm., σ_{dis} is 5.5 and σ_{ave} is 0.99. For the auscultatory method the average is 122.5mm., σ_{dis} is 5.85 and σ_{ave} is 1.05. The σ of the difference is 1.44. The reliability of the difference is 0.485σ .

In 23 of the 31 comparisons the readings were identical. The range of the difference is 2 to 6mm. This is very small as compared to the range of 8 to 25mm. obtained in the first series. The reliability shows that there is little or no difference between the mechanical and auscultatory methods and that they may be used interchangeably with only a very slight error which is negligible compared to the normal variations in lateral blood-pressure.

This is not the case with the palpation and mechanical

methods, however, for there exists a completely reliable difference between the two.

THE DETECTION OF DECEPTION BY INCREASES IN LATERAL BLOOD-PRESSURE

When this work started it was the intention of the investigator to use the Tycos Recording instrument throughout the whole group of experiments. However, it became necessary to return the sphygmomanometer to the Taylor Instrument Company before the series was completed and for the latter part of the work the Tycos Indicating instrument was employed with the auscultatory method.

The subjects used were undergraduates who were taking general psychology at Seth Low Junior College of Columbia University and whose ages range from sixteen to twenty-five years. Fifty-nine subjects were employed, the majority being male Hebrews.

PROCEDURE

All of the work was conducted in a research room in the Physics Building at Columbia University where excellent conditions for such an experiment may be obtained. The subjects were told before they came to the room that they were to undergo a deception test. The attempt was made at this point to set up in the subjects a desire to deceive the experimenter and to prove the test a failure. How well this attempt succeeded may be indicated by the fact that out of the first thirty-three tests only three subjects chose to tell the truth.

When the subjects came to the room they were given a card on which the following was written, "At half past nine Monday morning (date) the paymaster of the National Biscuit Company was attacked and robbed at the corner of Ninety-ninth Street and Broadway. He was hit on the head with a lead pipe and five thousand dollars were taken from him. You were seen at Ninety-eighth Street and Broadway at twenty-five minutes past nine. On Friday (date) your room was searched by detectives and three thousand dollars were found under the rug. Make up an alibi that will adequately cover the situation. Remember that you must admit those things which are known about you. In this test you

.

may assume either that you are guilty or not guilty. If you wish to assume that you are telling the truth, that you do not know anything about the hold-up, that you are not guilty, you will use the alibi furnished by the investigator. If you wish to assume that you are guilty, that you did commit the hold-up, that you will lie and try to get away with it, you will use the alibi that you have written. Whether you assume that you are guilty or not guilty you must write an alibi."

"Do not tell the investigator what you intend to do."

"Memorize the alibi that you intend to use for you will be examined."

By having each subject write an alibi, whether or not he intended to use it, the investigator always kept on hand one that he had not heard, which might be used for truth by the subject who followed.

The instructions were repeated orally by the investigator and the subject was asked if he thoroughly understood the instructions. The investigator then left the room while the subject wrote an alibi and memorized the one he chose to use. As soon as the subject was ready he called the investigator back to the room. The subject then rolled his right shirt sleeve to his shoulder and seated himself at a table with his back to the instrument. The band was then adjusted to his right arm and inflated. The questioning began. The subject was asked his name, age, residence, telephone number and occupation outside of school hours. While these questions were being asked and answered, one and in some cases two readings were taken. Although the subject was not instructed to answer truthfully to these questions he almost always did so. In those cases in which the subject lied in answering these questions the data were omitted. Without an appreciable pause the questioning was shifted to the alibi. No rigid form of questioning was followed. To some extent the questioning depended upon the content of the alibi. In general, such questions as these were asked:

"Where were you at 9:30 o'clock on Monday morning (date)?"

"What were you doing there?"

"Did you see the hold-up?"

"Were you connected with the hold-up?"

"Did you hit the man with a lead pipe?"

"Who did?"

"What did you do with the other \$2,000?"

"Why did you hide the \$3,000?"

"Is that a lie?"

During the time these questions were being asked and answered three records were taken. The interval between readings in this and all succeeding work was one minute.

RESULTS

The data taken for Truth are shown in Table III, those for Deception in Table IV. The first column gives the subject's initials; the second, third, fourth and fifth columns show the four readings taken on each subject. The sixth column shows the average of these. Figure IV shows the curves plotted from data in Tables III and IV.

Tables V and VI show the differences between the first and each of the three following readings. The sixth columns show the average and the seventh the greatest of these differences.

There are six ways of considering these data which are as follows:—

(1) By the use of the Absolute readings of the group.

Table III shows that the averages of the four readings for the truth group increased from 125mm. on the initial reading to 129.8mm. on the third, from which it drops to 128.5mm. The reliability of the difference between the average of first readings and that of each of the following is 0.18σ , 0.90σ and 0.69σ respectively. This means that there is an unreliable difference between that part of the test in which subjects were telling the actual truth and that in which they assumed themselves to be answering truthfully about an alleged crime.

Table IV shows that the averages of the four readings for the deception group increases steadily from 124mm. on the first reading to 143mm. on the fourth. The reliability of the differences between the average of the first and each of those following are 1.54σ , 3.7σ and 4.3σ respectively. This means that there is a completely reliable difference between readings taken during actual truth and those taken during the assumption of an attitude of deception.

(2) By the use of average readings.

This is the way in which Landis and Gullette treated their data and is the grossest possible.

The average truth reading is found by averaging the readings in the sixth column of Table III. This is 127.4mm. The σ is 13.4 and σ_{ave} 2.7. The average deception reading, found in the same way, is 134.3mm. The σ is 12.7 and σ_{ave} 2.2. The difference between the averages is 6.9mm. The σ_{diff} is (shown in Table III) 2.58 and the reliability of the difference is 2.46σ . There are 993 chances in 1000 that there is a difference between these groups. The figure obtained by Landis and Gullette for this was 820.

(3) By the use of differences.

Tables V and VI show the differences between the first reading and each of the following for Truth and Deception respectively. The sixth columns show the averages, the seventh the greatest of these differences.

In the truth group the difference between the first and each of the three following is 1, 5, and 3.5mm. respectively. The corresponding sigmas are 3.66, 5.8 and 8.1, and σ_{ave} is 0.752, 1.16 and 1.6.

In the deception group the difference between readings 1 and 2 is 7mm. The σ is 7.8 and σ_{ave} is 1.33. The difference between readings 1 and 3 is 15mm., with a σ of 8.9 and σ_{ave} of 1.53. The difference between readings 1 and 4 is 20mm., with a σ of 10.3 and σ_{ave} of 1.77. The σ_{diff} between the averages of the differences of similar readings in the Truth and Deception groups is found in Table VI. [$\sigma(\text{diff})$ (n&n)] as are the reliabilities of these differences. The reliability of the difference between the averages of the changes in the case of the first readings is indeterminate and is assumed to be 0.⁹ The reliability increases to 4σ , 5.2σ , and 6.9σ respectively in the following averages.

(4) By use of the Average Difference.

Instead of using the averages of the differences between the first and each of the three following readings the average of all the differences may be used. These are shown in the sixth columns of Tables V and VI. These show the average

⁹ That this assumption is warranted is shown by the very low reliability of the difference between the averages of initial readings in Tables III and IV.

change for Truth to be 2.3mm., with a σ of 3.6, a σ_{ave} of 0.72 and the average for Deception to be 10.5mm. with a σ of 5.8 and a σ_{ave} of 1.0. The reliability of the difference between these averages, shown in Table V, is 6.8σ .

(5) By use of the Greatest Difference.

If, instead of the Average Difference Columns we use the seventh columns of Tables V and VI, which show the greatest differences and treat them as in the preceding case, the reliability of the difference between the averages of the greatest differences in the two groups is 6.65σ .

(6) By means of Overlapping of the two groups.

In all of the above ways of treating the data we are working with the reliability of a difference between the two groups. These measures, no matter how great the $R(\text{diff})$ is found to be, tell nothing of the accuracy with which truth and deception are differentiated. This can be determined best by the overlapping of the greatest differences in the two groups. Figure I shows the greatest differences for each of the subjects of the Truth group, plotted above the axis and those for the subjects of the Deception group plotted below. This diagram shows that the greatest accuracy is obtained if 13mm. is taken as the critical pressure. Five subjects who assumed the attitude of deception show less and three who chose that of truth show more than the critical pressure. The accuracy of the differentiation is 87%.

Figure II shows the outstanding individual difference in the blood pressure curves obtained from the truth group and Figure III those from the deception group. In considering these, the outstanding feature is the wide range of variation. In deception the initial pressure ranges from 100 to 166mm. No two curves are identical or even very similar. There are 34 subjects in the deception group and, with very gross classification nine kinds of curves are found. This classification is according to slope alone and is irrespective of the general level of the curve.

Subjects AB and LO conform to curve 1 in which the level is fairly constant through the first three readings with a rise on the fourth. Subjects HC and JO showed results similar to curve 2 in which the high point comes in the 2nd reading followed by a decrease to a constant level. Subjects AN, WK, CZ, IR, EB, RC, IL, HL, AN, HSm, MH, HS, NP, EW and

FR gave curves somewhat similar to 3 in which there was an increase in each succeeding reading from the first to the fourth. Subjects BM, GM, and JE gave curves similar to 4 in which there was an alternating direction of slope. Subjects TK, IS and AH gave curves similar to 5 in which there is a rise to the 3rd point followed by a fall on the fourth. Subjects EF, MK, BMi, TL, NG gave curves similar to 6 in which the pressure was constant through the first 2 readings followed by a rise to the fourth. Curve 7 is plotted from the results of subject LB in which the pressure decreases from the first to a constant level. Subjects PF and GS gave curves similar to 8 in which there is a rise from the first to the 3rd reading where it becomes constant. Subjects ME and PK gave curves similar to 9 in which there is a fall in the 2nd followed by a rise in the 3rd and 4th readings.

If any of these curves were to be taken as typical it would probably be curve 3 to which 15 of the subjects conform more or less closely. In the light of the number of different kinds of curves obtained it would probably be better to say there is no typical deception curve unless we mean by that the curve drawn from the averages which is shown in Figure IV.

In the truth group Figure II we find much the same sort of a condition existing. Subjects AK, AS and AW had changes similar to those shown in curve 1 in which there is a slight rise in the 2nd and 3rd points with a decrease on the 4th. Subjects ASt, SG, Tn and GP had changes similar to that shown in curve 2 in which the pressure is fairly constant in the 1st and 2nd readings and rises on the 3rd and 4th. Subjects IR, KD, IK, BSm, HJ and BS gave results similar to curve 3 in which the pressure was fairly constant throughout. Subjects BS and HJ had changes similar to those shown in curve 4. Curve 4 was plotted on the results obtained with subject BS in which the pressure decreases to a constant level on the 3rd and falls on the 4th readings. Subjects IL, JK and WC had changes similar to those shown in curve 5 in which there is a rise from the 1st to a constant level in the 2nd, 3rd and 4th readings. Subjects VD and MG had changes similar to those shown in curve 6 in which the pressure remains constant to the 3rd and drops on the 4th reading. Subjects CK and JD had changes similar to those shown in curve

7 in which the slope alternates. Subjects JT, BJ and JS had changes similar to those shown in curve 8 in which the pressure rises from the 1st to the 4th readings. Curve 9 was plotted from data obtained from subject MR in which the pressure rose to the 2nd and fell to the 4th reading. Here again with gross classification we find 9 kinds of curves in the data obtained from twenty-five subjects.

In this group the initial pressure ranges from 104 to 152mm. It is evident that variation is the normal condition and that irrespective of blood pressure levels there is no really typical truth curve unless one chooses to say that a curve plotted from the group averages as shown in Figure IV is the typical truth curve.

CONTROL EXPERIMENTS

In the experiment described above a group curve for deception has been given. This curve gives rise to three questions:

- (1) Can it be produced in any situation other than that of deception?
- (2) Will deception give such a curve under all circumstances, and
- (3) What is the cause of the rise?

To answer these it was found necessary to devise three control experiments. In the first an attempt was made to find a situation in which there was no deception but which gave a curve similar to that obtained in deception. In the second the subjects chose an attitude of truth or deception in a situation which had little or no emotional content. A third experiment was found necessary as a control for the first. In this it was shown by elimination what had caused the rise in the first.

EXPERIMENT A

CHANGES OCCURRING DURING AN INTELLIGENCE TEST

In this experiment the subjects were presented with a situation which has a tendency to disturb somewhat the emotional equilibrium of most people. They were told that their intelligence was to be measured.

Throughout this work the Tycos Recording Sphygmomanometer was used to measure the blood-pressure. The "intelligence test" consisted of a multiplication sheet of two-place numbers. The eighteen subjects were a heterogeneous group all of whom were taking summer work at Columbia University. Six were instructors of psychology; one or two others were candidates for the degree of Ph. D. in psychology; the remainder were students who were taking courses in general psychology; four were women.

It was necessary to deceive the subjects to some extent in order to induce them to come to the laboratory. Few people enjoy having their intelligence measured. Therefore, they were lured to the room under the pretext of having their normal blood-pressure changes observed. This was not entirely untrue.

The subjects were seated as in the deception test with their arms resting on a table and their backs toward the instruments and the observer at their right side. The band was attached to the right arm and four readings of normal blood-pressure were taken. The investigator then said the following to each subject: "I wish to discover if there is any relation existing between the range of normal blood pressure changes and intelligence. To do this I shall now give you a battery of tests and take blood-pressure readings while you are working. The first is a test of mathematical ability. You will multiply the columns of numbers. Give your answers aloud. Work as rapidly as possible. There is a time limit and your score is computed in terms of the numbers which you complete correctly."

The phrase "battery of tests" was introduced only to make the procedure seem more forbidding. There was no test other than the multiplication sheet.

Many of the subjects objected to having such a test foisted upon them and said so in no uncertain terms. Some started to work and declared that they would not go further. A demonstration of disgust on the part of the investigator in these cases had the desired effect and the unruly subjects went on working.

Four readings were taken while the subjects worked. The interval between readings was about one minute.

RESULTS

The data taken in this experiment are shown in Table VII. The average changes for the group are very similar to those obtained in the deception test. There is a steady increase from the 1st to the 4 readings where the pressure remains about constant. The greatest difference is between the first and the fourth averages, the difference being 15.5mm. The average of the greatest individual changes is somewhat larger, 20.6mm. as we should expect.

These rises are not as great as those obtained in the deception test. Possibly one explanation of this is the difference in the subjects used. In this group, as we have noted before, are many people who have been working in psychology for some time. Probably these have been subjects for many psychological experiments. As a result they become so accustomed to the experimental situation that they show less emotional change than does the naive subject. If we omit these subjects from the group the averages become somewhat different. The initial pressure is 113.4mm. and rises to 135.3mm. on the fourth readings, an increase of 21.9mm. The average of the greatest individual changes is 26.5mm. (Fig. V.).

These changes are somewhat larger than those obtained in the deception tests. It is needless to draw conclusions as to the quantity of the change. It is merely pointed out that the quantitative differences between the averages obtained in this experiment and those obtained in the deception experiment are not important. The slope of the curve is the same throughout both experiments. That these results are so very similar to those obtained in the deception experiment is probably accidental. The range of individual difference is again very great. The range of change is from a fall of 5mm. in the case of subject MS to a rise of 35mm. in subjects CL and KH. Subject MS is a Ph. D. candidate with whom the investigator has been closely associated for the past two years. He was quite familiar with the work that was in progress and had been a subject in other parts of the work. He multiplies very rapidly and accurately and at the end of the test said, "You should have given me something difficult—I was once a bank clerk so this is very simple to me." Subject

RN became very excited and angry during the test and said, "I can't multiply" and completed the test only upon being urged. Subject KH became disturbed and said, "You had better give me something else to do." Subject MH refused to do any of the work as soon as she was told it was an intelligence test. It seems probable that in this case there was a desire to avoid effort.

The changes were probably not all due to the same emotional conditions as we have seen in some cases there was anger and in some cases what seemed to be more closely allied to fear. In other cases the condition was probably one of excitement of the variety found in competitive situations.

No matter what caused the rise the fact is that we have produced changes similar to those obtained in the deception experiment under conditions in which the subject did not lie or have an attitude of deception. We are, therefore, justified in concluding that lying is not characteristic of the so-called lying curve.

EXPERIMENT B

BLOOD-PRESSURE CHANGES DURING MENTAL WORK

In the preceding experiment it was shown that large rises were obtained due to some factor present in the situation in which the subject thought his intelligence was being measured. There are two factors which may have some connection with the rise. The subjects were somewhat tense or excited and they were also doing mental work. If we can eliminate one of these factors there may be some justification in assuming that the other is the cause of the change.

This experiment was designed to test the effects of mental work which is entirely free from any excitement. The necessity of reporting in such a situation is apt to lead to some changes that would not occur if the situation were entirely free from checks and competition.

The Tycos Recording and the Tycos Indicating Instruments were used in this experiment. The multiplication sheets used in the "intelligence test" were used again in this series.

The twenty-seven subjects were undergraduates enrolled in the courses in general psychology at Seth Low Junior Col-

lege of Columbia University. The age range is from sixteen to nineteen years. The nationality is predominantly Hebrew. Four of the subjects were girls. Eleven of them were tested in the research room of the Physics Building. The remainder were tested in a small office at Seth Low.

Each subject was given the following instructions orally: "I wish to find out what effect pure mental work has upon blood pressure. I shall give you a sheet upon which are lists of numbers to be multiplied without the aid of paper or pencil. You need not tell me the answers you get nor will you be asked how many you have completed at the end of the test. I have no check on how much you do. If you do nothing I shall not know it. All that I have is your assurance that you will work hard and complete as many as possible while you are being tested." Each subject promptly assured the experimenter that he would do as many as he could. The results obtained are based solely on the assumption that the subjects did work throughout the test. Many gave objective manifestations of effort but no record was kept of these.

The subjects were seated as before and five readings were taken, one before he started to work and four while he was working. The data taken are shown in Table VIII.

RESULTS

The averages obtained show a drop in the blood pressure from 123mm. on the first reading to 121mm. on the fifth. The pressure remained constant through the first three readings, the drop of 1mm. occurring in each of the fourth and fifth readings. The reliability of the difference between the averages of the first and fifth readings is very small. The σ_{diff} between one and five is 3.36; the difference is 2 and the reliability of the difference is 0.595.¹⁰

The range of variation in the change is again large from a drop of 14mm. in subject AB to a rise of 11mm. in subject MS. In eighteen of the twenty-seven subjects there was a decrease in pressure. This decrease may have been due to the fact that the subjects began the test with a very slight

¹⁰ These results agree with those of Marston on mental work. Binet and Vaschide, *L'Année Psy.* 1896, found that intense mental work caused a rise of 20m.m. Hg. Their subject reported his results as he completed each problem.

amount of excitement which disappeared as the test progressed.

The curve plotted from the averages is shown in Figure VII. The results show conclusively that the rise obtained in the intelligence test experiment was not due to the effort expended in mental work. It is then probably due to the other factor which we have noted formerly—that of agitation or excitement.

EXPERIMENT C

CHANGES DURING UNEMOTIONAL DECEPTION

We saw in Experiment A that the "lying curve" was found under conditions other than deception. It was further shown that a rise was due to excitement. Two questions remain to be answered: (1) What part of the rise found in the deception situation is due to excitement? and (2) How does the consciousness of lying influence the pressure? To answer these Experiment C was designed.

The Tycos Indicating instrument and the auscultatory method were used throughout this series. All readings were taken on the left arm.

The 45 subjects used were of the same type as those used in the experiment on mental work. Three were girls. Eight of the subjects were tested at Seth Low and the remainder in the Physics Building.

In this series there was no attempt to induce in the subjects any spirit of competition. Rather it was attempted to implant the impression that the deception was only incidental and of little interest to the investigator.

The subject was given a card upon which the following instructions were written: "You will be given a card on which there are listed ten objects. You will be asked the color of these objects. You must answer truthfully when asked the first five. You may answer either truthfully or falsely to the last five. Do not mix your answers. If you answer one of the last five falsely answer all five falsely. Choose whichever attitude you like. This is not a deception test."

Before being tested each subject was asked to repeat the instructions. He then chose one of two cards on which were listed ten objects. These two cards were the same except

that the last five objects on the cards were arranged in different orders. This difference in order prevented the experimenter from recognizing the attitude assumed by the subject. It would have been much easier for the experimenter if some mechanical means had been used to make all the subjects lie, such as drawing lots to determine which attitude he would assume. It is believed by the present experimenter that forcing an attitude on the subject in such a way introduces another factor, the results of which cannot be measured.

The subjects were seated as in the previous experiments with the instrument attached to the left arm. The questioning was as follows:—"What is the color of the first object?" "What is the color of the second object?", etc. As in the other experiments the interval between readings was about one minute. In this case five readings were taken, two while the subject was responding truthfully and three while he was answering truthfully or falsely, according to the attitude which he had chosen. At the close of the series the subject was asked what attitude he had assumed.

RESULTS

Table IX shows the data taken from the truth group; Table X those of the deception group and Table XI a summary of these. Figure VI shows the curves plotted from the averages taken from Table XI.

The curve for deception starts at 136mm. and remains practically constant to the fourth reading. Between the fourth and fifth readings there is a drop to a final pressure of 133mm.

The truth curve starts at a pressure of about 129mm., slowly increases to a little over 130mm. on the third reading from which it drops to about 129 on the fifth reading.

The reliability of the difference between the initial averages is 1.97σ . This reliability decreases to 0.89σ between the final averages.

The shapes of these curves are very different from those obtained in the first experiment. Probably in this case not even the difference in levels is of any significance since by omitting one subject who consistently shows a high pressure the general level of the deception curve may be lowered

about 2mm. The difference between the two is probably due for most part to chance.

The range of changes in the truth series is from a fall of 14mm. to a rise of 6mm. The range in the deception group is from a fall of 24mm. to a rise of 10mm. Only three subjects in this group showed a rise of over 4mm.

From these results it is evident that deception in itself, free from other exciting factors, does not cause the lateral blood-pressure to rise. The rise that was obtained in the first experiment must have been due to excitement. It is further evident that there is no characteristic curve for deception even when group results are used.

SUMMARY AND CONCLUSIONS

A. We have shown in the introduction that the major difference between the work of Marston and of Landis is in the treatment of their data. Marston considered individual changes and Landis worked with the grossest of averages which blanketed any individual changes. Further, Landis' own results do not justify his conclusion.

B. In a preliminary investigation the relations between the three methods of obtaining blood-pressure were determined. The mechanical and the auscultatory methods were found to be equally accurate. The palpation method was found to be on an average of 17.6mm. less accurate for this investigator than the mechanical method. Since the mechanical and auscultatory methods are equally accurate about the same difference exists between the auscultatory and palpation methods.

C. In the deception experiment the subject was presented with a laboratory situation in which he might either lie or tell the truth. Using lateral blood-pressure as the criterion, truth and deception were differentiated with 87% accuracy. The curves obtained from individuals conform to no particular type in either truth or deception. Variation is the outstanding characteristic. However, group curves for truth and deception were drawn which show a completely reliable difference. The average rise was found to be 5.1mm. for truth and 20.8mm. for deception.

D. In the first control experiment the subject was pre-

sented with a situation which he believed to be an intelligence test. The curve obtained from the group averages was very similar to that found in the deception experiment. In this case there was no deception. Again, there is no typical curve and the range of variation is wide.

E. In the second control experiment the subject was presented with the same mental work that was performed in the first. The conditions under which he worked were designed to eliminate excitement. The curve obtained was practically a horizontal line indicating that the rises found in the second experiment were due to excitement and not to mental work .

F. In the third control experiment the subject was presented with a situation in which he might lie or tell the truth but which, unlike that of the first deception experiment, was free from exciting conditions. The curves obtained show no rise in lateral blood-pressure. Both the truth and deception curves remained practically constant in level throughout and in no way resemble those previously obtained.

From these results the following conclusions may be drawn:

- 1st. Lateral blood-pressure may be used to detect deception experimentally when the deception situation gives rise to excitement and when other causes of excitement are eliminated.
- 2nd. The "deception curve" may be obtained under conditions other than those of deception, as when subjects are told that their intelligence is being measured.
- 3rd. Mental work causes no rise in the blood-pressure and, therefore, the rise obtained in the intelligence situation is due to excitement.
- 4th. The consciousness of deception causes no change in lateral blood-pressure. The rise obtained in the first experiment is, therefore, due to excitement.
- 5th. There is no characteristic curve for deception.

COMPARISON OF BLOOD PRESSURE METHODS

Table I
MECHANICAL versus
PALPATION
Readings of Lateral Pressure

<i>Machine</i>	<i>Palpation</i>	<i>Difference</i>
128	115	13
126	109	17
134	110	24
129	112	17
132	110	22
136	115	21
130	105	25
125	106	19
127	109	18
129	104	25
136	115	21
127	105	22
128	110	18
127	107	20
120	107	13
129	115	14
125	115	10
123	115	8
116	102	14
123	105	18
123	103	20
127	105	22
125	105	20
120	108	12
117	108	9
117	100	17
115	100	15
120	100	20
112	97	15
Av	125 107 4	17 6
σ m	=5 77	
σ (ave) m	=1 15	
σ p	=5 08	
σ (ave) p	=0 945	
σ (diff) (m & p)	=1 43	

Reliability of the difference between
(m) and (p) 12 3 σ

Table II
MECHANICAL versus
AUSCULTATORY
Readings of Lateral Pressure

<i>Machine</i>	<i>Ausculta</i>	<i>Difference</i>
120	120	0
120	120	0
126	124	2
124	122	2
124	122	2
118	118	0
120	118	2
119	115	4
121	121	0
123	121	2
120	120	0
121	121	0
121	121	0
122	122	0
114	114	0
124	124	0
117	117	0
122	122	0
115	115	0
120	120	0
119	117	2
121	121	0
127	127	0
127	212	6
126	126	0
127	127	0
125	125	0
125	125	0
132	132	0
140	140	0
140	140	0
Av.	123 2 122 5	0 71

σ m =5 5
 σ (ave) m =0 99
 σ a =5 85
 σ (ave) a =1 05
 σ (diff) (a & m) =1 44

Reliability of the difference between
(m) and (a) 0 485 σ

Table III

Experiment 1

DETECTION OF DECEPTION

Data for Truth Group—25 subjects

Readings of Lateral Blood Pressure in mm. Hg.

<i>Subject</i>	<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>Average Reading</i>
B S	140	137	137	129	136
J T	112	113	130	132	122
B J	125	128	132	137	130 5
A St	141	137	148	152	144 5
J D	120	114	118	111	115 7
V D	120	120	121	110	117 7
J S	112	115	117	121	116 3
T N	114	114	117	124	117 3
H J	134	130	129	132	131 3
K D	109	106	107	108	107 5
A W	120	119	135	133	126 7
I K	128	128	128	130	128 5
I L	114	124	124	124	121 5
B Sm	106	106	110	106	107
A S	146	150	154	150	150
J K	112	116	120	120	117
A. K	152	158	169	148	156 7
S G	126	126	134	136	130 5
W C	134	140	140	144	139 5
C K	116	120	116	124	119
G. P	118	118	124	130	122 5
I. R	140	140	140	140	140
B C.	134	136	136	132	134 5
M G.	148	148	152	140	147
M. R.	104	110	108	104	106 5
Ave	125	126	129 8	128 5	127 4
σ	13 7	13 8	13 6	13 3	13 4
σ (ave)	2 74	2 76	2 72	2 66	2 7
r (1 & n)		95	91	79	
σ (diff) (1 & n)		5 4	5 3	5 1	
R (diff) (1 & n)		0 18 σ	0 90 σ	0 69 σ	
σ (diff) between average of Average Reading (T & D)—2					58
R (diff) between average of Average Reading (T & D)—2					46 σ

Table IV

Experiment 1.

DETECTION OF DECEPTION

Data for Deception Group—34 subjects

Readings of Lateral Blood Pressure in mm Hg

<i>Subject</i>	<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>Average Reading</i>
W K	113	130	137	140	130
E F	118	118	134	151	130
M K	100	98	113	116	107
B M ₁	134	132	153	158	144
C Z	110	116	121	129	119
I R	122	134	146	150	138
B. M.	138	148	140	168	148 5
E B	123	124	138	142	132
R C	120	122	126	144	128
I L	115	132	143	150	135
P F	115	125	146	146	133
T K	122	125	145	140	133
M E	115	110	122	128	119
H L	116	124	125	130	124
A N	130	155	160	165	152
H Sm	116	126	130	131	126
I S	117	130	130	129	125
G M	108	116	125	122	118
M H	119	130	142	146	129
H S	139	145	146	151	145
G S	127	134	145	145	138
T L	135	132	156	162	146
P K	110	104	122	123	115
N P	127	131	146	153	134
J E	136	151	150	171	152
E W.	120	132	140	153	137 5
J O	120	144	142	142	134
F R.	116	135	145	151	137
N G	134	134	140	148	139
H C	150	162	156	156	156
A. B	166	166	164	174	168
A H	142	150	160	150	151
L. O.	124	122	126	126	124
L. B.	118	116	116	116	116 5
Ave.	124	131	139	143	134 3
σ	12 9	15 2	13 0	14 8	12 7
σ (ave)	2 22	2 58	2 23	2 52	2 2
r (1 & n)		78	76	75	
σ (diff) (1 & n)		4 53	4 05	4 44	
R (diff) (1 & n)		1 54 σ	3 7 σ	4 3 σ	

DIFFERENCES BETWEEN THE FIRST AND EACH OF THE
FOLLOWING READINGS FOR TRUTH

	<i>Difference Between First and Each of the Following Readings</i>					
<i>Subject</i>	<i>1st</i>	<i>2nd</i>	<i>Srd</i>	<i>4th</i>	<i>Average Difference</i>	<i>Greatest Difference</i>
B. S.	0	- 3	- 3	-11	-4	-11
J T.	0	1	18	20	9·7	20
B J.	0	3	7	12	5·5	12
A. St.	0	- 4	7	9	3	11
J D.	0	- 6	- 2	- 9	-4·3	- 9
V D.	0	0	1	-10	-2·3	-10
J S.	0	3	5	9	4·3	9
T. N.	0	0	3	10	3·3	12
H. J.	0	- 4	- 5	- 2	-2·7	- 5
K D.	0	- 3	- 2	- 1	-1·5	- 3
A. W.	0	- 1	15	13	6·7	15
I K.	0	0	0	2	5	2
L L.	0	10	10	10	7·5	10
B Sm.	0	0	4	0	1	4
A S.	0	4	8	4	4	8
J K.	0	4	8	8	5	8
A. K.	0	6	17	- 4	4·7	17
S G.	0	0	8	10	4·5	10
W C.	0	6	6	10	5·5	10
C K.	0	4	0	8	3	8
C P.	0	0	6	12	4·5	12
I R.	0	0	0	0	0	0
B C.	0	2	2	- 2	5	2
M G.	0	0	4	- 8	3·0	- 8
M R.	0	6	4	0	2·5	6
Av.		1	5	3·5	2·3	5·1
σ		3·66	5·8	8·1	3·6	8·8
σ (ave.)		752	1·16	1·6	0·72	1·7
R (diff) between averages of Average Difference (T. & D.)					-6·8 σ	
R (diff) between averages of Greatest Difference (T. & D)					-6·65 σ	

Table VI

DIFFERENCES BETWEEN THE FIRST AND EACH OF THE
FOLLOWING READINGS FOR DECEPTION GROUP

		<i>Difference Between First and Each of the Following Readings</i>				<i>Average Difference</i>	<i>Greatest Difference</i>
<i>Subject</i>		<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>		
W K		0	17	24	27	17	27
E F		0	0	16	33	12	33
M K		0	— 2	13	16	6 5	16
B M _i		0	— 2	19	24	10	24
C Z		0	6	11	19	9	19
I R		0	12	24	28	16	28
B. M.		0	10	2	30	10 5	30
E B		0	1	15	19	9	19
R C		0	2	4	24	8	24
I L		0	11	28	35	20	35
P F		0	10	31	31	18	31
T K		0	3	23	18	11	23
M E		0	— 5	7	13	4	13
H L		0	8	9	14	8	14
A. N		0	25	30	35	22	35
H S _m		0	10	14	15	10	15
I S		0	13	13	12	8	13
G M		0	8	17	14	10	17
M H		0	11	23	27	10	27
H S		0	6	7	12	6	12
G S		0	7	18	18	11	18
T. L		0	— 3	21	27	11	27
P. K		0	— 6	12	13	5	13
N. P		0	6	19	26	12	26
J E		0	15	14	35	16	35
E W		0	12	20	33	16	33
J O		0	24	22	22	17	24
F R		0	19	29	35	21	35
N G		0	0	6	14	5	14
H C		0	12	6	6	6	12
A B		0	0	— 2	8	2	8
A H.		0	8	18	8	9	18
L O		0	— 2	2	2	0	2
L. B		0	— 2	— 2	— 2	1 5	— 2
Av		0	7	15	20	10 5	20 8
σ		0	7 8	8 9	10 3	5 8	9 3
σ (ave)	0		1 33	1 53	1 77	1 0	1 6
σ (diff) (n & n)			1 52	1 92	2 4		
R (diff) (n & n)			4 σ	5 2 σ	6 9 σ		

Table VII

Experiment A

INTELLIGENCE TEST

18 Subjects

Readings of Lateral Blood Pressure in mm Hg

<i>Subject</i>	<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>5th</i>	<i>Greatest Change</i>
H L	119	121	142	122	124	23
J W.	123	126	121	127	133	10
D E	119	120	144	134	132	25
M S	100	95	96	97	98	-5
A R	126	130	141	128	136	15
R D.	118	118	122	137	134	19
W. K.	111	114	122	126	123	15
K T.	130	125	155	150	139	25
Y. Y.	131	131	157	153	151	26
C. L	112	124	128	147	139	35
E L	100	124	115	114	116	24
L G	89	104	113	122	118	33
M Y.	110	116	133	125	128	23
C D	135	152	140	150	156	21
M. S.	115	116	124	110	110	9
M H.	125	120	130	128	126	5
K H	117	112	129	143	152	35
R N	95	104	104	129	108	34
Ave.	115	119.5	128	130.5	129	21.5
σ	12.6	11.8	15.8	15.3	15.1	10.7

Table VIII

Experiment B

MENTAL WORK

27 Subjects

		Readings of Lateral Blood Pressure in mm Hg.					
<i>Subject</i>		<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>5th</i>	<i>Greatest Change</i>
M S		109	120	115	110	110	11
R N		110	104	108	111	111	— 6
A W		125	126	125	121	121	— 4
J D		116	118	120	121	121	5
J N		132	132	131	131	131	— 1
T N		117	125	125	115	115	8
J S		110	99	103	105	105	—11
B F.		114	112	112	110	110	— 4
A M		130	126	130	130	124	— 6
I H		116	116	114	116	114	— 2
E. F.		142	148	148	144	136	6
P A		144	146	150	146	136	— 6
F M		114	114	114	114	114	0
M G		136	136	144	138	142	8
W B		124	124	124	122	120	— 4
W R		124	124	118	120	118	— 6
X. L		114	116	116	114	110	— 4
A B		146	134	148	132	138	—14
J S		100	90	90	94	100	—10
A H		130	126	126	128	128	— 4
S G		112	118	116	112	118	6
M. K		150	150	150	148	146	— 4
L S		116	114	118	114	116	— 2
J K		108	110	104	104	106	— 4
M R.		114	108	108	106	106	— 8
S D		128	134	130	130	128	2
W. C.		130	130	126	126	124	— 6
Av.		123	123	123	122	121	—2 15
σ		12 7	13 6	14 8	13 4	12 0	6 25

Table IX

Experiment C

UNEMOTIONAL LYING

Data for the Truth Group—24 Subjects

		Readings of Lateral Blood Pressure in mm Hg					Greatest Change
<i>Subject</i>		<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>5th</i>	
A	H	142	150	148	148	150	2
H	C	142	140	140	140	138	— 2
I	K	120	126	126	132	126	6
A	B	160	164	166	166	166	6
B	S	106	118	116	110	104	12
A	S	142	136	142	146	146	— 6
M.	R	92	88	92	88	86	— 6
J	A	120	124	124	126	120	6
S	D	146	138	138	140	138	— 8
W.	C	140	142	140	143	140	3
I	R	148	148	148	146	140	— 8
F	G	130	136	136	138	136	8
X	X	110	114	119	112	116	9
S	M	142	142	146	146	152	10
B	C	132	132	132	130	130	— 2
J	H	130	134	132	136	132	6
T	M	124	126	126	128	126	4
C	B	124	118	128	118	114	—10
C.	M.	130	128	128	124	128	— 6
J	W.	134	130	138	134	136	— 4
W	B	124	130	124	130	126	6
F	M.	110	110	120	114	118	8
I.	H	118	114	118	118	116	— 4
B.	F	118	120	120	118	118	2

Table X

Data for the Deception Group—21 Subjects

		Readings of Lateral Blood Pressure in mm. Hg.					
<i>Subject</i>		<i>1st</i>	<i>2nd</i>	<i>3rd</i>	<i>4th</i>	<i>5th</i>	<i>Greatest Change</i>
M	P	118	114	114	116	108	—10
I	L	122	130	128	132	122	— 8
S.	G	122	123	126	130	124	8
S	C	126	122	126	120	120	— 6
D	C	152	148	150	148	150	— 4
J	K	130	128	138	138	128	8
L	S	132	138	136	138	136	6
N	G	138	122	122	128	128	—16
P	M	134	140	142	138	142	8
J	Kl.	130	128	130	128	128	— 2
A	S	146	138	140	141	142	— 8
Z	T	128	120	128	126	126	— 8
D	G	142	138	140	134	134	— 8
H	I	138	142	144	140	138	6
A	G	150	150	150	150	140	—10
M	F	136	136	132	138	140	4
M	G	140	142	142	140	138	2
W	R.	120	120	120	130	124	10
P	A	154	158	160	154	154	6
D	F	156	168	164	150	140	28
A	M	138	138	134	134	134	— 4

Table XI

RELIABILITY OF THE RESULTS OBTAINED IN THE
CONTROL DECEPTION TEST

<i>Reading</i>	<i>Truth</i>			<i>Deception</i>			<i>Reliability</i>		
	<i>Av</i>	$\sigma(dis)$	$\sigma(ave)$	<i>Av</i>	$\sigma(dis)$	$\sigma(ave)$	$\sigma(diff)$	<i>R</i>	<i>Chances</i>
1	128 5	15 2	3 10	136	19 1	2 21	3 80	1 97 σ	1000
2	129 5	15 3	3 12	136 6	13 4	2 92	4 26	1 67 σ	975
3	130 5	15 8	3 22	136	10 7	2 33	3 98	1 38 σ	916
4	130	13 6	2 78	136	9 5	2 08	3 46	1 73 σ	958
5	129 3	17 1	3 50	133	10 5	2 29	4 17	0 89 σ	813

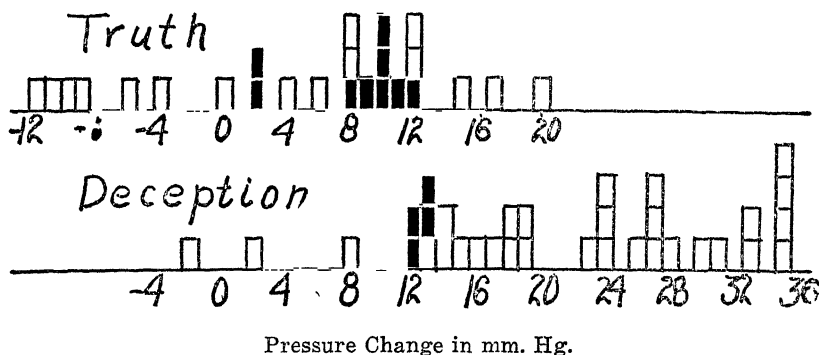


Figure 1.—Distribution of Changes in Truth and Deception.

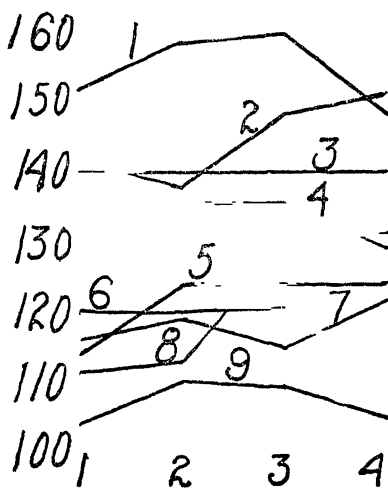


Figure 2.—Curves Showing Individual Differences in the Truth Group.

1. A.K, AS, AW; 2. AS_t, SG, TN, GP; 3. IR, KD, IK, BS_m, BC, HJ; 4. BS; 5. IL, JK, WC; 6. VD, MG; 7. CK, JD; 8. JT, BJ, JS; 9. MR.

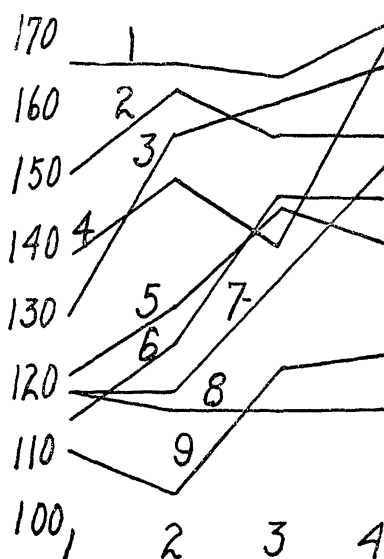


Figure 3.—Curves Showing Individual Differences in the Deception Group.

1. AB, LO; 2. HC, JO; 3. AN, WK, CZ, IR, EB, RC, IL, HL, HS_m, MH, HS, NP, EW; 4. BM, GM, JE, FR; 5. TK, IS, AH; 6. PF, GS; 7. EF, MK, BM, TL, NG; 8. LB; 9. ME, PK.

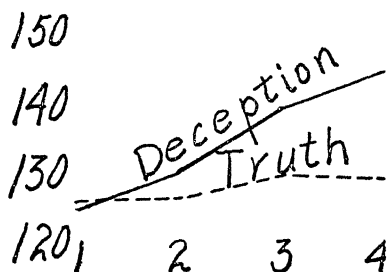


Figure 4.—Group Curves for Truth and Deception.

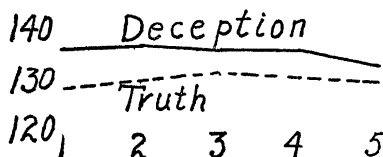


Figure 6.—Group Curves for Unemotional Truth and Deception.

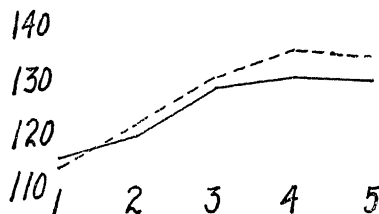


Figure 5.—Group Curve for "Intelligence Test."

Total Group is indicated in the full line.

Group with Ph.D. candidates omitted is indicated by the broken line.

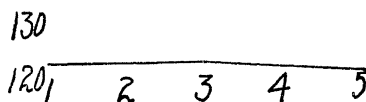


Figure 7.—Group Curve for Mental Work.

VITA.

Matthew N. Chappell, born in Wakefield, Rhode Island, July 26th, 1900. Bachelor of Science in Electrical Engineering from Rhode Island State College in 1924. Columbia University Scholar 1927-1928, Columbia University Summer School Scholar, 1928. Sigma Xi, 1928. Instructor of Psychology in Columbia University Extension 1927-1928. Instructor of Psychology in Columbia University 1928-1929.

THE MEASUREMENT OF VERBAL AND NUMERICAL ABILITIES

BY
MATTHEW M. R. SCHNECK

Submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy, in the Faculty of Philosophy,
and Pure Science, Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, Editor

No. 107

NEW YORK
June, 1929

ACKNOWLEDGMENTS

For suggestion and definition of the problem, for the general plan of the investigation, and for almost daily consultation and guidance with respect to procedure and statistical treatment, I am deeply indebted to Professor Henry E. Garrett, of the Department of Psychology, Columbia University. Without Professor Garrett's direction, this study could not have been made. To the Department of Psychology, Columbia University, I extend thanks for relieving me of the task of calculating many correlations. The Department was kind enough to finance this work, which was done by the Columbia University Statistical Bureau. I am also indebted to Professor Howard D. Marsh of the Department of Philosophy, College of the City of New York, for permission to use his classes in Experimental Psychology, and for valuable cooperation in other directions. I am grateful for the kindness of Professor Egbert Turner, who turned over to me his classes in Educational Psychology at the College of the City of New York, for preliminary experimentation. My thanks are also extended to Mrs. Georgette Decsi Schneck for valuable assistance in the scoring of the tests and in calculations.

THE AUTHOR

TABLE OF CONTENTS

<i>Chapter</i>	<i>Page</i>
I. The Problem	5
II. The Literature	7
III. The Procedure	15
1. The Subjects	15
2. The Tests	15
3. The Testing	18
4. The Scoring	20
IV. The Results	21
1. Means, Standard Deviations and Intercorrelations	21
2. Statistical Analysis and Discussion	23
3. Relation of Verbal and Numerical Abilities to Other Abilities	40
4. Relation of Verbal and Numerical Abilities to Scholastic Records	42
V. Summary	48
VI. Bibliography	49

The Measurement of Verbal and Numerical Abilities

I. THE PROBLEM

This paper is a specific contribution to an increasing body of literature dealing with the organization of mental traits. The center of interest for the present study lies in the interrelationships existing between two manifestations of intellectual function, the term 'intellect' being used in much the same sense as that in which it is used by Thorndike (22). These functions may be termed, for want of more precise names, *verbal ability* and *numerical ability*. The writer hesitates to assign names of any kind to these activities, for the use of substantives to denote activities is always hazardous, and almost certain to give rise to needless controversy. The designations will not be insisted upon, nor will they be defended. It is desirable, nevertheless, to explain what is meant by the terms employed. For the purpose of this paper, then, verbal ability will refer simply to performance in those tests which involve, predominantly, a knowledge of words and ability to use them. The essence of the performance demanded by those tests which will be called verbal is the possession by the subject of the accomplishments just mentioned. Similarly, numerical ability will refer to performance in those tests which involve, predominantly, the ability to manipulate number and number concepts. Verbal and numerical tests are designated as such in the body of this report.

It is not pretended that exhaustive tests of either kind of ability have been made, nor that complete measurements of 'verbality as such' or of 'numerical ability as such' have been achieved. An attempt was made, indeed, to secure a fairly representative sampling of these two abilities. Whether or not the attempt has succeeded is a matter which the writer prefers to submit to the judgment of the reader. This, however, is insisted upon. Whatever the abilities involved in the tests may be, it is with these abilities that the writer is concerned. Since definitions in psychology are sometimes less than definitive, it is felt to be better merely to designate rather than to define.

If, then, it be permitted to refer to the ability or abilities called upon in the verbal tests here employed as *verbal ability*, and to the ability or abilities involved in the numerical tests as *numerical ability*, the following questions will be dealt with:

1. Are verbal ability and numerical ability discrete and independent capacities, or do they possess elements in common?
2. Are there group factors of verbal ability and of numerical ability, either or both?*
3. Are measurements of these abilities merely measurements of ability in general?
4. What is the best single test or battery of tests for the measurement of these abilities?
5. What is the relationship between verbal and numerical ability, on the one hand, and, on the other hand, performance in college courses presumably involving these abilities?

* In the system of Spearman, a group of abilities may be divided into a single general factor which is common to all of the abilities, plus specific factors, each of which is present in one, but not more than one, ability. This is true provided a certain statistical criterion, to be described later, is satisfied. If specific factors overlap, so that they are present in more than one, but not in all of the abilities, the criterion cannot be satisfied, and a group factor, caused by the overlapping, is present.

II. THE LITERATURE

A search for literature bearing directly upon the foregoing problems yields scant return. There have been, indeed, a host of measurements of verbal ability and of numerical ability, but few of these studies have gone beyond a mere calculation of coefficients of correlation, from the magnitude of which extremely dubious conclusions have been drawn. The popular notion has been that verbal intelligence is something quite different from mathematical intelligence, and that an individual may be proficient in the one, and, at the same time, deficient in the other. This notion gains some support from a study of the correlations reported. These are nearly always positive, but rarely high. Thorndike (23) has said that "The ability measured by verbal tests is not the same as the ability measured by non-verbal tests." This seems to be the prevailing notion, though Spearman holds a contrary view. According to Spearman (17) all mental performances partake of one general common function called 'g', and both verbal and non-verbal tests really measure this function.

C. Burt (2, p. 46ff) administered a series of tests to 120 children "of the same sex, differing but little in age, zeal, attendance and social status," and found the following special abilities: arithmetical, manual, linguistic and composition. The linguistic group, according to Burt, is not entirely unrelated to the composition group, and the two may perhaps be taken as forming a single literary group. Burt's method of locating these special abilities is of doubtful accuracy, depending as it did upon an inspection of his correlation tables. He observed that the values found in this table did, to a certain extent, conform to a theoretical hierarchical arrangement, which indicated the presence of a common general factor. In some cases, however, the correlations are far too high; and these cases he takes to be indicative of the presence of special factors additional to the universal factor common to all of the subjects. Now, the method of inspection of a table to determine the presence or absence of hierarchy is decried by Spearman himself (17, p. 138). (For the significance of the hierarchy, see Spearman, (17)). A better approximation is obtained by the use of the intercolumnar correlation coefficient (17), but even this has been rejected by Spearman in favor of the cri-

terion of the tetrad difference. (17, p. 79). It may be added that the present writer's inspection of Burt's correlation table yields conclusions which are by no means in agreement with those of Burt. It appears to the writer that no particular arrangement of any kind is present, and that neither a universal factor nor special factors can be inferred with safety.

A more recent study is that of Davey (6). As a pupil of Spearman, in whose laboratory the problem of group factors is a very live issue, Miss Davey undertook the comparison of ability in verbal tests with ability in tests which were of similar form, but non-verbal. She constructed a battery of pictorial tests similar in form to a series of oral tests. Her subjects were 243 children in London County Council schools, ranging in age from 8 to 14 years, and divided into ten groups. Six of these groups were in different classes of different schools. A seventh group consisted of 82 boys, all in the same school, divided into 4 sub-groups according to age. The total number of subjects of each sex is not given, nor is the number of subjects in each of the groups. It is obvious, however, that her groups must have been very small. There was no mixture of the sexes in any of the groups, but there may have been diversity in age, excepting in the 4 sub-groups of boys, mentioned above. Intercorrelations among the tests given were calculated for each of the ten groups, and these intercorrelations averaged. The average intercorrelations seem to have a satisfactory degree of reliability, but it is regrettable that the individual coefficients were not published. An average of ten correlations may be itself reliable when the single correlations, or some of them, are unreliable. Working with such small groups, it is very probable that some of the correlations were unreliable. The propriety of working with such average correlations may be questioned.

Eight oral and six pictorial tests were given to each of the groups, and the correlations calculated by means of the Spearman Rank-order formula. Averaging the intercorrelations, it appears that the oral and pictorial tests correlate to about the same extent as do the pictorial tests among themselves, but that the correlation for the oral tests *inter se* is somewhat higher. The P. E. of the difference in the latter case (calculated by the present writer) is not completely significant, if 4 P. E. be accepted as the significant criterion. Miss Davey, of course, does not accept these data as definitive evidence of

anything. She calculated 420 tetrad differences, and found that the median tetrad difference is .021, with a probable error of .0194, which she interprets as indicative of the presence of a group factor. This view is rather surprising, coming from Spearman's laboratory, for in the latter's volume, *The Abilities of Man*, a median tetrad difference which is not much larger than its probable error is uniformly regarded as insignificant.

Analyzing the table of tetrad differences, Miss Davey concludes that there is no group factor in the pictorial tests; that there is a group factor in the oral tests; but that this group factor is confined to the first four of these oral tests. From this result, she comes to the further conclusion that there is no group factor of verballity as such, for if there were, it would have shown itself in all eight tests. Just what is meant by 'verballity as such' is not made entirely clear, though some light is thrown upon the matter by a further analysis, as the result of which Miss Davey concludes that the group factor was introduced by the content of the tests, rather than by their form. Even if we grant the validity of the method by which this conclusion was arrived at (and it is by no means granted), there still appears to be a distinction without much difference. Form as distinguished from content is altogether a matter of interpretation. The distinction made here seems to be another attempt to reify an abstract concept. To the present writer, this situation seems to justify the notion of Wilson (24) that 'g' is relative to the set-up, unless one can get a situation in which $r_{ag} = 1$ *. It is probably true that suitable changes in the content of Miss Davey's first four oral tests might have eliminated the group factor which she found. This would prove merely that it is possible to devise tests which are so very different from one another that no group factor is present. It would prove nothing about 'verballity as such'. Granted that the form of the tests is not responsible for the introduction of the group factor; granted that the responsibility lies with the content, or with the similarity of the relations educed; granted further, that a change in content would eliminate the group factor; we must necessarily conclude that, since the form does not change, content is responsible both for the group factor when it is present, and for its absence when it is absent. All of which seems to reduce form, so far as group factors are

* r_{ag} = the correlation between the general factor, 'g' and any given ability, 'a'.

concerned, to insignificance. Verbality as such seems to be dependent upon the set-up, and to have no meaning whatever apart from the set-up. Miss Davey's final conclusions are to the effect that a verbal test measures the same 'g' as does a test similar in form but non-verbal. The conclusion may express a truth, but it certainly does not follow from Miss Davey's results. At least, her own verbal tests measure some group factor which is absent from her pictorial tests.

A. I. Gates (10) used verbal and non-verbal material for the purpose of predicting success in school subjects. Twenty third-grade subjects were given seven non-verbal and five verbal tests. The average intercorrelations follow:

Verbal	.62
Non-verbal	.40
Verbal and non-verbal	.24

From these results Gates concludes that the verbal tests clearly yield results different from the non-verbal. While it is unsafe to draw conclusions from data obtained from so small a group, the results point in the direction of the prevailing opinion.

T. L. Kelley (15), working with students in the seventh and third grades and in the kindergarten, has located group factors as follows: verbal, number, memory, spatial and speed. He has, in addition, devised a method for measuring these group factors. In the same volume (p. 192ff) Kelley analyses the data reported in Rose G. Anderson's monograph (1). Rearranging and combining Anderson's results, and applying the earlier steps of his own procedure, Kelley locates a verbal, a numerical and a memory bond.

In general, a verbal factor seems to be indicated in all of the studies cited. By reason of the criticisms advanced above, the evidence appears to be less than conclusive. Even Kelley admits that his groups were not to his liking so far as homogeneity is concerned. This matter of homogeneity of groups is a vexing one. Groups composed of individuals differing materially from one another in age, nurture, nationality, social and economic conditions, and the like, are dangerous to work with, for the correlations obtained may be spurious. Thus, if a group of boys and girls, in equal number, and of the same age, be measured for strength of grip and for general intelligence, and the correlations between these capacities be cal-

culated, it is obvious that the correlation will be spurious. The boys will be high, mediocre and low in intelligence, as will, also, the girls. But the boys, as a body, will rank higher in strength of grip than will the girls. In such a case, the correlation between strength of grip and general intelligence is very likely to be zero, or even negative. The same condition will hold if the correlation were calculated between general intelligence and height for a group composed of Japanese and Swedes. The correlation will certainly be lower than it would be if the group were composed entirely either of Japanese or of Swedes. Again, Kelley (15, p. 25) has shown how a correlation may be raised by the introduction of diversity in maturity. So that diversity within a group may raise a correlation, or it may lower it.

The distorting influence of diversity has been stressed by Kelley (15) and by Spearman (17, p. 155), the former insisting that heterogeneity tends to introduce correlation, the latter contending that heterogeneity will reduce it. Either result may follow, as has been shown above. In view of Spearman's insistence upon homogeneity, it is curious that two of the principal studies upon which he relies for the maintenance of his thesis were, in this respect, glaring offenders. Thus, Spearman makes much of the study of Bonser (17, pp. 139, 147, xii, xx). But Bonser's group of 757 subjects was composed of 385 boys and 372 girls, ranging in age from 8 to 16, in the 4th, 5th and 6th grades of five public schools. This is heterogeneity with a vengeance. Again, Spearman stresses the study of Simpson (17, p. 145). For these data, Spearman calculated 3003 tetrad differences, and hails the result as one of the most striking agreements between theory and observation ever found in psychology, adding that this agreement would not easily be matched in any other science. But if ever a group was heterogeneous, it is precisely the group upon which these momentous conclusions are based. Simpson worked with a total of 37 subjects, which is obviously a very small group. Worse than this, the subjects were 17 advanced students and professors, and 20 adults of the sort which may almost be termed, appropriately, human derelicts. It is difficult to imagine a greater degree of diversity than just this group. Even if it be agreed, with Spearman, that heterogeneity reduces correlation, it is surely a strange procedure to make use of a widely diverse group merely because correlation happens

to be present. It may very well be that the correlation is present just because the group is heterogeneous. There will be occasion to advert to this matter later.

In addition to the verbal factor which has been indicated, the studies of Burt, Kelley and Anderson point to the existence of a numerical factor. The evidence, on the whole, is meagre. A few studies do publish correlations between arithmetic and verbal tests. Thus, Bonser (4), working with children in the fourth, fifth and sixth grades, finds a correlation of $+.41$ between arithmetic and completion tests, and a correlation between arithmetic and opposites amounting to $+.42$. Whether these data do or do not indicate special factors is not determined. Indirect evidence of more or less independence of traits is furnished by studies of special disabilities. Bronner, for example, finds marked disability for language in cases where all other processes are normal. She reports similar clinical findings for numerical ability. (5, Chaps. 5 and 6). Special disabilities of this sort are also reported by Symonds (19), Hollingworth and Winford (13), C. Burt (3) and A. I. Gates (11).

The problem of group factors has assumed major importance in recent years, largely as a result of the assiduous labors of Spearman and his pupils. One study from that laboratory has already been reviewed above. In general, Spearman (17, Chap. 13) is not greatly impressed by the evidence for group factors. After analyzing this evidence, he states that "cases of specific correlations or group factors have been astonishingly rare." He does find evidence for group factors "in respect of what may be called the logical, the mechanical, the psychological, and the arithmetical abilities." He acknowledges also a special ability for appreciation of music. A slight degree of confusion appears in Spearman's account, for he employs the term "group factor" in more than one sense. Thus, he recognizes perseveration, oscillation and persistence of motives as universal factors, but in several places he also refers to them as group factors. The latter, in Spearman's system, are defined as "those which occur in more than one but less than all of any given set of abilities." It is in this sense that the writer will speak of group factors in this paper.

Since Spearman's statistical technique is, to a certain extent, employed in this report, it seems desirable to devote some space to it. It is not deemed necessary to trace the development

of the technique, since this has already been done by Dodd (7, 8). Briefly, then, Spearman holds that when certain statistical conditions are satisfied, it is possible to divide cognitive abilities into two parts, independent of each other. The one part is a *general* factor, called 'g'. It varies from individual to individual, but remains the same for a given individual in respect of all of the correlated abilities. The other part is *specific*, varying from individual to individual, and, for any given individual, from ability to ability. The statistical condition which must be satisfied is as follows:

If a group of individuals is tested in a number of abilities, the intercorrelations between all possible pairs of the variables will, if arranged in the usual correlation table, furnish the necessary data for the application of the method. If we let 'a' stand for one variable, 'b' for a second, and 'p' and 'q' for a third and fourth, it can be shown that the abilities tested are divisible into one general factor and four specific factors when

$$r_{ap} \times r_{bq} - r_{aq} \times r_{bp} = 0$$

throughout the table. The entire expression is called the tetrad equation, and the right side of the equation is termed the tetrad difference. When four variables are used, three tetrad equations are calculated. With a greater number of variables, the number of tetrad equations to be calculated is given by $3_n C_4$. Fortuitous influences are always present, so that the tetrad equation will rarely equal zero exactly. It must be shown, then, that the probable error of the tetrad difference is large enough to make the difference insignificant. When this can be shown for all of the tetrad differences in the table, then the abilities in question can be divided as above. When tetrad differences are significantly greater than 0, we have evidence of the presence of a group factor or factors which Spearman interprets as the overlapping of specific factors. The proof of these propositions is given in (17, Appendix, iii).

The system sketched above has been attacked very vigorously, notably in a series of papers by G. H. Thomson (21), who presents some trenchant criticisms of Spearman's interpretations, especially of 'g'. The present writer is, in this paper, very little concerned with the nature of 'g'. For his purposes, the divergent views of both Spearman and Thomson are quite immaterial. The tetrad criterion does indicate the

presence of something which is shared by all of the variables, whatever that may be. Further than this one need not venture, though the writer will have occasion to offer some remarks bearing upon this central factor.

III. THE PROCEDURE

1. *The Subjects.* Reference was made above to the desirability of working with a group of subjects in which diversity is reduced in so far as this is possible. The writer was fortunate in having, in his own classes, such a group. All of the subjects used were students in a class in General Psychology, at the College of the City of New York. The tests were given to the entire class, and the unsuitable subjects eliminated after the testing. Those retained numbered 210. All of them were men of Hebrew extraction, whose parents were of foreign birth. The group was further limited to those whose parents were born in Russian, Polish or Balkan countries. A few of the subjects were sons of professional men, but these were eliminated, so that the final group included only those whose fathers were engaged in manual trades or in small retail establishments of their own management. Seven upper sophomores and lower seniors are included, the remaining 203 subjects being in the junior year. All of the subjects were educated in the New York City public schools, including the elementary school, the high school and the college, the latter being a municipal institution, tuition free. Nearly all were self-supporting, in whole or in part. In age, the subjects ranged from 18 to 21. This introduces diversity, but diversity which is less damaging than would be a similar age range in the earlier years. Furthermore, age was partialled out before any of the significant statistical steps were taken, so that in this respect the group was homogeneous. To summarize: The group is fairly large, and quite similar with respect to education, ancestry, economic and social status and age. All were of the male sex. On the whole, heterogeneity was reduced to a minimum.

2. *The Tests:* Nine tests in all were given to all of the subjects. Five of these were intended to measure verbal ability, and four were intended to measure numerical ability. The tests were of the following types:

A. *Verbal Tests.**

1. *Vocabulary Test:* This contained 130 items, of the selective type, in which the person tested was required to choose,

* The writer will be glad to supply copies of the tests to interested persons.

from a group of four words, that one whose meaning was most nearly the same as that of the test word. The instructions were as follows:

"Look at the first word in Line 1. Find the other word in the line which means the same or most nearly the same as the first word. Underscore this synonymous word.

helotry diabolism gambling slavery assistance

The word which is most nearly the same as helotry is slavery, which is therefore underscored. Do the same in each line below. Underscore one word in each line and only one."

About 30 words were taken from the CAVD test and about ten from the Detroit Advanced Intelligence Test, as well as a few from the Ohio State University Intelligence Test. In most cases, the writer found it necessary to make changes in the words from which choice was to be made, in order to fit the test for a group of upperclass college men. The greater bulk of the test was composed by the writer. The net result was a test of unusual difficulty, but not too difficult for this group.

2. *Opposites Test*: This test was of the same form and length as the Vocabulary Test, and of equal difficulty. The instructions were similar, except that the testees were required to underscore the word which was most nearly the opposite of the key word. A few of the words in this test were taken from the Ohio State University Intelligence Test, the rest being prepared by the writer.

3. *Analogies Test*: This consisted of 40 items, ranging from easy to very difficult items. The instructions were as follows:

"In each of the lines below you will observe the relationship between the first and second words. From among the words contained in the parentheses, choose the word which bears the same relation to the third word of the line as the second word bears to the first. Thus:

Sky : blue :: grass : (table green warm big)

Here the word green, contained within the parentheses, bears the same relationship to 'grass' as 'blue' does to 'sky'. The word 'grass' is therefore underlined. Do the same in each of the lines below. Underline one word in each line, and only one word. Remember: The word to be underlined is always in the parentheses." Eighteen of the analogies were taken from Army Alpha and from the Miller Mental Ability Test. The remainder, comprising the more difficult items, were prepared by the writer.

4. *Sentence Completion Test*: This was composed of 50 items, all taken from the CAVD Test, Levels M, N, O, P, Q. The instructions given were those used in the published test.

5. *Disarranged Sentences*: This test contained 40 items, all prepared by the writer. Following are the instructions given:

"Read the sentences as they would be if the words were arranged in correct order. Do not rewrite the sentences. When the sentence asks a question, answer it on the blank line. When no question is asked, do what the sentence tells you to do."

B. *Numerical Tests*:

1. *Arithmetic Reasoning Test*: This test was composed of 40 items, of various types. The items were taken from the CAVD Test; from the Otis Self-Administering Test of Mental Ability, Higher Examination, Form A; from Morgan's Mental Test, Form A; from the Brown University Psychological Examination, Series II, Exercise F; and from the Rogers Test of Mathematical Ability. A few were prepared by the writer.

2. *Number Series Completion*: This test contained 40 items. Four were taken from Army Alpha, the remainder being composed by the writer. The instructions were:

"Look at each row of numbers below. On the two dotted lines write the two numbers that should come next, as in the samples."

Samples:

2	4	6	8	10	12	14	16
9	8	7	6	5	4	3	2
						---	---

3. *Equation Relations*: This test contained 26 items, of two forms. Sixteen of the items were of the type given in the CAVD Test, some of them being drawn directly from that test, the rest being added by the writer. The instructions for these problems were:

"In the problems below, write the numbers and signs in the proper order, so that they make a true equation. The letters in the problem stand for arithmetical signs, as shown in the key. Whenever a letter appears, refer to the key and convert the letter into the proper mathematical sign."

Key: A stands for the plus sign.

B stands for the minus sign.

C stands for the multiplication sign.

D stands for the division sign.

E stands for the equals sign.

F stands for parentheses, i.e., ().

Samples: Problem

3 3 6 E A
4 7 8 20 E A C

Solution

$3 + 3 = 6$
 $7 \times 4 = 20 + 8$

Ten items were of the form given below. The instructions follow:

"In the problems below, only the equality sign is given. It is indicated by the letter E, as above. You are to supply the signs needed, and to write the numbers and signs in the correct order, so that they make a true equation. The numbers which you find at the left of the E will remain on that side in the finished equation. Those at the right of the E will remain on the right. You are simply to rearrange the numbers on each side of the E, and to supply the missing signs, to make a true equation."

Thus: 2 7 E 3 18 3 Solution: $7 + 2 = 18 - (3 \times 3)$

These items were all prepared by the writer. They proved to be somewhat less difficult than the items of the other form.

4. *Mental Multiplication:* This test contained 58 items, composed by the writer. The instructions follow:

"In each row you will find a multiplicand and a multiplier, as indicated by the column headings. Multiply *mentally* the multiplicand by its accompanying multiplier, and write the result in the column marked 'Product.' Thus, in the first row, the multiplicand is 14; the multiplier is 11. Multiply the 14 by the 11 and write the result on the dotted line to the right of the 11. Do the same for each row. *Remember:* do no figuring with pencil and paper. All calculations must be done mentally, without any aids."

It is acknowledged that the tests could have been improved. This is particularly true of the test of Disarranged Sentences, which proved to have been somewhat too easy. Scores for this test were piled up at the upper end of the curve. The Analogies Test should have been longer, for its reliability turned out to be somewhat lower than was desired. In general, the tests were fairly satisfactory, so far as reliability is concerned. (See Results.) Only the Analogies and the Disarranged Sentences failed to meet the standard set, which was a reliability of .90 or better. The reliability of the Analogies was .879, and the Disarranged Sentences only .755.

3. *The Testing:*

All of the tests were administered by the group method. The subjects were given mimeographed sheets which contained the

test material. The Analogies, Disarranged Sentences and Opposite tests were given in two groups of 120 each. The remainder of the tests were given in the laboratory, to groups of 18. An attempt was made to have the items of the tests arranged in order of difficulty. This order was determined by preliminary testing of students in classes in Educational Psychology. None of these students were in the experimental group itself. It was found that the order of difficulty determined in this way was maintained fairly well in the experimental group. Difficulty was determined by the number of failures for each item.

With the exception of the test in Mental Multiplication, all tests were power tests. Multiplication was a speed test. The time allowed was as follows:

Vocabulary	35 minutes
Opposites	35 minutes
Analogies	30 minutes
Disarranged Sentences	25 minutes
Sentence Completion	50 minutes
Arithmetic Reasoning	45 minutes
Number Series	30 minutes
Equation Relations	50 minutes
Multiplication	10 minutes

The total testing time was, therefore, *5 hours and ten minutes*. The testing was distributed over three weeks, it being thought undesirable to harass the students unnecessarily. It is believed that the subjects entered into the program in a spirit of cooperation. The tests were sufficiently difficult to constitute a genuine intellectual challenge. In the case of the student-body of the College of the City of New York, this is all that is necessary to insure interest and enthusiasm. Few of the students had ever been tested before, so that it was not found necessary to deal with a condition of negative adaptation to a testing program. The work was done at a time when the intelligence-testing movement was under discussion in class, and the program, therefore, fitted naturally into the general scheme of things. The members of the class were not told the purpose of the experiment, and most of them did not realize that an experiment was going forward. They were told, however, that the results of the testing would have no effect upon their term grades. In general, then, the unnatural conditions of experimentation were minimized. The time allotted for the tests proved to be sufficient, so that 90% of the subjects were

able to finish the work, except, of course, in the case of the Multiplication. Some of the men did finish this test in the ten minutes allowed, but most did not. None, however, failed to complete at least 30 items in Multiplication.

IV. *The Scoring*: All tests excepting the Sentence Completion were scored in terms of the number of items done correctly. The sentences were scored on a scale ranging from 0 to 3, the latter being allowed for completions admitted as correct by the key used for the CAVD Test at Teachers' College, Columbia University. The score of 0 was given in cases of complete failure, or when the completion did not make good sense. The scores of 1 and 2 were allowed for cases in which the completion did not agree with the key, but in which there was not failure. When only one word differed from the key, the score of 2 was given. When more than one differed, the item was allowed one point. In both cases, of course, the sentence as completed had to make reasonably good sense. Since this scheme was followed for all subjects, it is felt that the results of scoring by this method were valid.

The Disarranged Sentences were so constructed that only one correct response was possible. It was observed that, in nearly all cases, the response was either correct or there was no response at all. One item of the Arithmetic Test admitted of two correct answers, and both were accepted. This was the first item in the test. In the Number Series Completion Test, it was required that both dotted lines be filled in correctly. If either was incorrect, the whole item was counted as incorrect. The Equation Relations and the Multiplication presented no difficulties. In the Multiplication, the answer was either correct or it was not. In the Equations Relations, only those equations which were correct throughout were accepted. No partial credit was given. The correct responses in the Vocabulary, Opposites and the Analogies Tests were determined by reference to manuals of synonyms and antonyms, and the words from which selection was to be made were so chosen as to permit only one correct response.

In general, then, the scoring was done with a high degree of objectivity. Subjective judgment was, in some degree, inevitable. Actually, individual judgment is never absent altogether, even in physical measurements. It is less prominent in some situations than in others, but it cannot be wholly avoided.

IV. THE RESULTS

1. *Means, Standard Deviations and Intercorrelations*

Means, standard deviations and intercorrelation coefficients were computed by the Columbia University Statistical Bureau, all coefficients being calculated by machine and verified carefully. The reliability coefficients were calculated by the writer. The odd-numbered items were correlated against the even-numbered items, and the result corrected by the application of the Spearman Brown Prophecy Formula (9, p. 269, Formula No. 59), to give the reliability of a test of double length. The means, standard deviations and reliability coefficients are given in Table I. The reliability coefficients will be found printed in italics, and lying along the diagonal from the first to the tenth variable. The tenth variable is age, whose correlation with each of the tests was calculated, age being stated in months. The inclusion of age as a variable was deemed desirable for the purpose of determining its effect upon the results, by the use of the partial correlation technique.

It will be observed that in all cases except one, that of Mental Multiplication, the correlation with age is negative. This is quite consistent with the usual findings. Within a given grade, the younger students are frequently of superior ability, and this seems to be true in the junior year in college. The negative correlations are low, but they seem to be fairly uniform in all cases excepting Multiplication. Here the correlation is positive and very slightly above zero. Actually, it may be taken as an absence of correlation. The reason for this exceptional value for Multiplication and Age is not entirely clear. It may be conjectured that, in such a relatively mechanical operation as multiplication, growth in ability approaches its limit at or before the age of 18, so that there is little differentiation to be found as the result of maturity or of precocity. It may be that the superiority of the younger members of a grade or college year is evidenced in respect of the relatively involved mental operations, those requiring greater reasoning ability and relation-perceiving ability, than in respect of the relatively easy or mechanical operations. If this be true, the absence of negative correlation would be explained, for multiplication makes less demand upon "intellectual" pro-

TABLE I
INTERCORRELATIONS AND RELIABILITY COEFFICIENTS OF NINE TESTS AND OF AGE, TOGETHER
WITH THE MEANS AND S.D.'S OF THE TESTS

	1	2	3	4	5	6	7	8	9	10
1. Vocabulary	.906	.8818	.6878	.5362	.3020	.2711	.1827	.0476	.0575	-.2377
2. Opposites		.925	.7170	.5374	.2646	.2889	.1778	.0539	.0458	-.2081
3. Analogies			.879	.4496	.2857	.2478	.1660	.0540	.0356	-.1727
4. Sentence Completion				.921	.2573	.3619	.3966	.1806	.1140	-.2883
5. Disarranged Sentences					.755	.1161	.0403	.0701	-.0255	-.2166
6. Arithmetic Reasoning						.920	.4517	.3754	.3728	-.2236
7. Number Series Completion							.907	.2512	.3083	-.2399
8. Equation Relations								.923	.2702	-.2497
9. Multiplication									.954	.0087
10. Age										1.0000
Mean	73.85	80.20	27.37	64.69	32.87	21.21	31.29	13.53	38.62	235.0
S.D.	15.74	15.15	4.18	20.07	3.95	6.25	6.73	3.70	11.16	11.3

Note: The Mean and S.D. of Age are given in months.

cesses than do the operations involved in the other tests. The writer does not insist upon this explanation, but he does regard it as probable, ruling out, of course, cases of phenomenal arithmetical ability as well as those cases in which some special scheme of mental multiplication is employed. Such schemes do exist, and the writer has no means of knowing how many of his subjects availed themselves of these short-cuts. A short-cut of this type would certainly tend to reduce even further the 'intellectual' elements in multiplication.

The reliability coefficients are, on the whole, satisfactory. All of the numerical tests have a reliability of .90 or more. Two of the verbal tests fall below this standard, but even in these cases the reliability is fairly high. The reason for the deficiencies of these tests has been suggested above.

It is possible to obtain a first impression of the relation between the verbal and the numerical tests by an inspection of the raw coefficients. The average intercorrelations are as follows:

Verbal Tests	.4920
Numerical Tests	.3383
Verbal & Numerical Tests	.1441

These results may be compared with those of Gates, cited in the foregoing. Gates, also, found that the average intercorrelations for the verbal and non-verbal tests was lower than those for either the verbal or non-verbal *inter se*. The evidence at this point is not definitive, but already it begins to point to a substantial difference between the verbal and numerical tests. The analysis must, of course, be carried much farther before final conclusions may be drawn.

2. Statistical Analysis and Discussion

A. It would have been desirable to confine the experimentation to subjects who were all of the same age. Unfortunately, this was not possible, for the size of the group would have been reduced quite drastically. Since age was included as a tenth variable, and its correlation with each of the other variables computed, it becomes possible to eliminate the influence of age by statistical treatment. If, then, age be partialled out, we have remaining the intercorrelations of the tests with the age factor held constant. This was done, with the results shown in Table II. Since the correlations with Age were, in all cases but one, negative, the effect of holding Age constant is to re-

duce the intercorrelations. Only where Multiplication is involved are the correlations raised, since the correlation of this variable with Age was positive.

B. The next step involves the correction of the coefficients in Table II for chance variations or errors of observation. Each of the coefficients was therefore corrected for attenuation, using the formula applicable when only one test for each variable is available. (9, Formula No. 48.) These corrected coefficients will be employed hereafter in the calculation of tetrad differences. According to Spearman (17, Appendix vi), the tetrad difference criterion is free from the influence of attenuation. "If the criterion is passed when the correlations are corrected for attenuation, then it must also be passed when they are not so." The writer does not dispute the validity of Spearman's mathematical demonstration of this proposition, *when the tetrad difference equals zero*. But tetrad differences rarely equal zero exactly. Spearman regards such a difference as insignificant when the tetrad difference is less than four or five times its probable error. It is quite possible, however, that a difference which, by Spearman's rather exacting standards, is insignificant, may become quite significant when corrected for attenuation. Conversely, it is conceivable that a tetrad difference which seems to be significant will become insignificant when the coefficients are corrected.

A coefficient of correlation is corrected for attenuation, by the method used in this study, by dividing the coefficient by the square root of the product of the reliability coefficients involved. Obviously, when the right side of the equation is zero, this division will not alter it. When, as is nearly always true, the tetrad equation does not equal zero the right side must be divided in the same manner as the left side, and by the same value. Now, the magnitude of the correction for attenuation depends upon the value of the reliability coefficients. The greater the value of the reliability coefficients, the smaller will be the corrections. If, then, the corrections to the right side of the minus sign in the tetrad equation be greater than those at the left of the minus sign, the tetrad difference becomes smaller than it would be in the absence of correction. If, on the other hand, the corrections to the left of the minus sign be greater, the tetrad difference will be larger than it would be without correction for attenuation. In this study, two tetrad differences which were statistically significant before correc-

TABLE II
INTERCORRELATIONS OF NINE VARIABLES WITH AGE PARTIALLED OUT

[illegible]

tion became insignificant after correction. It seems necessary, therefore, to correct coefficients for attenuation before the application of the tetrad criterion. Conclusions, otherwise, may be seriously erroneous. The corrected values of the coefficients of Table II are given in Table III.

These 'true' coefficients seem to indicate even more clearly than did the uncorrected data, the presence of a real difference between the verbal and the numerical tests. The average correlations follow:

Verbal	.5238
Numerical	.3507
Verbal and Numerical	.1223

Both for the verbal tests *inter se* and for the numerical tests *inter se*, the average intercorrelation is higher than before correction for attenuation. On the other hand, the average intercorrelation between the verbal and the numerical tests is lower. These data tell us nothing concerning the presence or absence of factors peculiar to either the verbal or the numerical abilities. Whether these factors exist must be determined by other means. We proceed, therefore, to the application of Spearman's technique, in so far as this technique is useful for present purposes.

C. The Tetrad Difference Criterion

When a correlation table contains nine variables, the number of tetrad differences to be calculated is 378. If each of these differences is zero, within the limits of its probable error, then each of the nine variables may be thought of as containing one general factor, shared by all of the variables, plus a specific factor or factors peculiar to each of the variables. The present problem, may, however, be approached somewhat differently. If our correlation table be divided into two parts, the one containing only the intercorrelations of the verbal tests, and the other only the intercorrelations of the numerical tests, we may calculate the tetrad differences for each of these two tables. It will then be possible to compute the correlation between the central 'verbal' factor and the central 'numerical' factor, provided the tetrad difference criterion be satisfied in both tables. This procedure eliminates much of the mechanical labor entailed by the calculation of 378 tetrad differences.

It will be observed (Table III) that the correlation between

Vocabulary and Opposites is .9569. This correlation is very much higher than are any of the others in the table, so high, indeed, that the two tests may be taken to involve identical abilities. This conclusion is supported by reference to the tests themselves. They are of the same form and of nearly equal difficulty. The only difference in the abilities called for is that in the one a relation of similarity is to be observed, while in the other the relation is that of opposites. It was decided, therefore, to eliminate one of these tests. An alternative to elimination would have been Spearman's procedure of pooling the results of the two tests. Since the judgment concerning the essential similarity of the two abilities is, in part, subjective, it was felt that the sounder procedure was the one decided upon. Either test might have been eliminated, but it was decided to eliminate Opposites, since the tetrad differences obtained by its inclusion were somewhat higher than when the Vocabulary test was used. It should be remarked, however, that in both cases the tetrad differences satisfy the criterion, as may be seen from the data given in Table IV. In this table, tetrad differences are designated in the manner adopted by Kelley (15, p. 11).

TABLE IV
TETRAD DIFFERENCES OBTAINED FROM TABLE III

<i>A. Verbal Tests</i>			
<i>Opposites Eliminated</i>		<i>Vocabulary Eliminated</i>	
	T.D. P.E.	T.D. P.E.	
t_{1234}	.0156 ± .0308	.0217 ± .0323	
t_{1243}	.0395 ± .0281	.0671 ± .0287	
t_{1342}	.0239 ± .0174	.0454 ± .0162	
<i>B. Numerical Tests</i>			
	T.D. P.E.		
t_{1234}	.0118 ± .0238		
t_{1243}	.0468 ± .0226		
t_{1342}	.0350 ± .0202		

In no case is the tetrad difference more than three times its probable error. The probable errors were calculated by the use of Kelley's formula* for the probable error of a tetrad

$$\begin{aligned}
 *P.E.t_{1234} = & .6745 \times \frac{1}{\sqrt{N}} [r_{12}^2 + r_{13}^2 + r_{24}^2 + r_{34}^2 + 2r_{12}r_{14}r_{23}r_{34} \\
 & + 2r_{13}r_{14}r_{23}r_{24} - 2r_{12}r_{13}r_{23} - 2r_{12}r_{14}r_{24} \\
 & - 2r_{13}r_{14}r_{34} - 2r_{23}r_{24}r_{34} + t_{1234}^2 (r_{12}^2 + r_{13}^2 + r_{14}^2 \\
 & + r_{23}^2 + r_{24}^2 + r_{34}^2 - 4)]^{1/2}
 \end{aligned}$$

difference (15, p. 49, Formula 28). This formula differs from the one offered by Spearman and Holzinger (17, Appendix, xi). Kelley points out the fact that Spearman and Holzinger's formula applies when the true tetrad difference is zero, whereas his own formula does not depend upon this. When the true difference is zero, then Kelley's formula becomes identical with that of Spearman and Holzinger. Holzinger (14) remarks that Kelley has retained a term which, in the actual calculation is negligible. This is true, but it was nevertheless deemed safer to use the longer formula, especially since the added labor is not irksome when only a few tetrad differences are calculated. Since each of our tables contains only four variables, it was not necessary to compute more than three tetrad differences for each. If, then, we accept the standard which requires that the tetrad difference be at least four times its probable error to be considered significant, it appears that the criterion is satisfied. It is probable that this standard is too exacting, but it seems best, at this point, to meet Professor Spearman on his own ground. It is concluded, then, that our verbal tests may be thought of as consisting of a central factor which is present in all of the verbal tests, plus a factor or factors specific to each of the tests, and not shared by any of the others. A similar conclusion is drawn for the numerical tests. The central factor for the verbal tests will hereafter be referred to as *V*, and that for the numerical tests as *N*. In neither case is any hypothesis advanced concerning the nature of these central or common factors.

D. Correlations between V and N

Since the major purpose of this study is the determination of the relation of verbal abilities to numerical abilities, it seems desirable to compute the correlation between *V* and *N*. Obviously, the specific factors involved have no place in this scheme since, by their very specificity, their correlations are zero, within the limits of their probable errors. If, then, there is any correlation between verbal and numerical abilities, it must consist in the correlation of their common 'general' factor or factors. The calculation of this correlation is arrived at by a circuitous route, to be detailed in the following pages.

1. The first step is the determination of the correlation between each of the verbal tests and *V*, as well as the correlation of each of the numerical tests with *N*. For this purpose, Spear-

man's formula (17, Appendix, Formula 21) was employed, yielding the results shown in Table V.

TABLE V
CORRELATIONS OF NINE TESTS WITH THEIR CENTRAL FACTORS, V OR N

	<i>Opposites Eliminated</i>	<i>Vocabulary Eliminated</i>
<i>A. Verbal Tests</i>		
Vocabulary and V	.8965	
Opposites and V		.8781
Analogies and V	.8160	.8435
Sentence Completion and V	.6081	.6128
Disarranged Sentences and V	.3841	.3622
<i>B. Numerical Tests</i>		
Arithmetic Reasoning and N		.7666
Number Series Completion and N		.5593
Equation Relations and N		.4613
Multiplication and N		.4867

The correlations with V differ somewhat, depending upon whether Vocabulary or Opposites be retained. These differences for the last three verbal tests should be well within the limits of the probable errors of these differences. These probable errors, then, were calculated, with the following results:

TABLE VI
PROBABLE ERRORS OF THE DIFFERENCE OF THE CORRELATIONS OF THE VERBAL TESTS WITH V AS DEPENDENT UPON THE ELIMINATION OF EITHER THE VOCABULARY OR OPPOSITES TESTS

	<i>Obtained Difference</i>	<i>P.E. of Difference</i>
Analogies and V	.0275	.0205
Sentence Completion and V	.0047	.0413
Disarranged Sentences and V	.0219	.0566

In no case, then, is the difference significant, as may be seen by a comparison of each of the obtained differences with its probable error.

The magnitude of the correlation coefficient of Table V is an index of the extent to which the several tests measure the common factors. The Vocabulary Test appears to be the best instrument for the measurement of that which is common to all of the verbal tests. Since the correlations have been corrected for attenuation, the reliability of the Vocabulary Test is 1.00. Its correlation with V is virtually .90, which is not far below its correlation with itself. Using uncorrected coefficients, the correlation of this test with V is .84, while its

obtained reliability is .91. A test whose correlation with a criterion is not far below its correlation with itself is a very satisfactory instrument for the measurement of that criterion. It should be recalled, in this connection, that Professor Terman (20) uses a vocabulary test in his revision of the Binet Scale as a preliminary determinant of the lower limit of the testee's mental age. He finds a knowledge of the meaning of words the best single criterion for this purpose. This is indirect evidence that 'general intelligence' as measured by the Stanford Binet Scale and other tests is largely verbal ability, or the ability to manipulate language.

The Analogies Test is somewhat less satisfactory as a measure of V, though its correlation with V is still above .80. This might have been expected, since this test also involves knowledge of words in much the same sense as does the Vocabulary Test. That it contains other elements, however, is indicated by the fact that the corrected correlation of Analogies with Vocabulary is only .76; and this fact leads to the conclusion that there is something more in V than mere ability to select synonyms.

The test in Sentence Completion is much less satisfactory than are Vocabulary and Analogies, while Disarranged Sentences seem to be distinctly unsatisfactory. It is the least reliable of the tests, and its distribution is noticeably skewed. Whether the shortcomings of these two tests lie in their form or in their content is not determined. Certainly their forms are quite different from those of the Vocabulary and Analogies tests. Their content must naturally be different, otherwise we would have a series of identical tests. Whatever these tests measure, they assuredly do not measure the same thing in equal degree.

None of the numerical tests is quite as satisfactory as is the Vocabulary Test or the Analogies. The highest correlation is that between Arithmetic Reasoning and N. Its magnitude is .7666. The other numerical tests are distinctly inferior to Arithmetic, so far as efficiency of measurement of N is concerned. No single member of our battery of numerical tests can be used as an approximately sufficient index of N. Whatever the nature of N, it appears that a considerable portion of it is left untouched by our numerical tests. This is true even when the entire battery is correlated with N, as will be shown presently.

2. It is possible to calculate the correlation of V with any combination of the verbal tests, to determine which battery will give the best measure of V. This can be done, also, for N and the numerical tests. These correlations are arrived at by finding the necessary multiple correlation coefficients. Thus, the expression $R_{1\ 2345}$ denotes the correlation between one variable and four other variables taken as a team. If, then, the tests be designated by numbers as shown in the following table, the multiple correlations may be found.

TABLE VII
CORRELATION OF V AND N WITH EVERY POSSIBLE COMBINATION OF TESTS
See Key Below

	<i>Verbal</i>	<i>Numerical</i>
$R_{1\ 2345}$.9306	.8294
$R_{1\ 234}$.9281	.8209
$R_{1\ 235}$.9240	.8205
$R_{1\ 245}$.9119	.8060
$R_{1\ 345}$.8590	.6953
$R_{1\ 23}$.9207	.8011
$R_{1\ 24}$.9070	.7906
$R_{1\ 25}$.9023	.7895
$R_{1\ 34}$.8545	.6574
$R_{1\ 35}$.8266	.6412
$R_{1\ 45}$.6533	.5833
Key: 1 = V & N		
2 = Vocabulary & Arithmetic		
3 = Analogies & No. Ser. Completion		
4 = Sent. Compl. & Equations		
5 = Dis. Sent. & Multiplication		

It is obvious that the best measure of both V and N is obtained with the use of the full batteries. When the last two tests of the verbal battery are used alone, the result is far from satisfactory. This is consistent with the data of Table V, where it was found that the correlation of the central factor with these tests, when taken singly, is relatively low. In the verbal set, the multiple correlation is reduced, whenever the Vocabulary Test is excluded, to a greater degree than when only the Analogies Test is omitted. This, again, is in accord with Table V. The same tendencies are observable in the numerical tests. Omission of the Arithmetic Test from any combination causes a sharp drop in the size of the multiple correlation. This test, taken alone, had the highest correlation with N. The numerical battery, taken as a unit, does not measure its central factor as efficiently as the verbal battery

measures its central factor. The numerical battery is better than the best single numerical test when taken alone, but it leaves something to be desired.

Whatever is common to the numerical tests is, in each of the single tests, apparently diluted by extraneous factors which are washed out when the group is taken as a unit. The nature of these extraneous factors is, at present, open to speculation. The Arithmetic Test includes the use of verbal elements, which may operate to reduce the correlation of this test with the 'numerical factor.' The other numerical tests also demand some verbal operations, but to a lesser degree than does the Arithmetic. It would seem, then, that the other tests ought to have a higher correspondence with N than does Arithmetic. The opposite situation, is, however, true. It may be, then, that the arithmetic test, which has a greater variety of operations than do the others, does by this very fact of variety include more of 'numerical ability.' Taken alone, then, the efficiency of this test is reduced by verbal factors. The addition of the other tests may result in the contribution of other 'numerical elements' which have the effect of offsetting the extraneous elements. They may, in addition, bring to bear other 'numerical elements' which are not present in the Arithmetic Test, over and above those which are required to nullify the effects of the verbal elements. It is admitted that this hypothetical explanation is a sheer guess, one which the writer does not care to defend very vigorously.

3. *Correlations of Tests with Specific Factors*

The mathematical derivation of the tetrad difference criterion is predicated upon the theory of partial correlation (17, Appendix iii). In this derivation, it follows from the definition of r_{ap} that $r_{ap_g} = 0$.* This means that when the tetrad difference criterion is satisfied, there can be no correlation between the specific factors contained in the variables. Such correlation as exists between the variables must be due entirely to 'g', the general factor which is shared by all of the variables. Since it is possible to find the correlation of each of the variables with 'g,' it must also be possible to find the correlation of each variable with its own specific factor or factors. This is

* R_{ap} designates the correlation between two variables, 'a' and 'p.' R_{ap_g} is the correlation between 'a' and 'p' with the factor which is common to both eliminated or held constant. R_{ap_g} , then, is the correlation between the factors which are specific to 'a' and 'p,' respectively.

accomplished by determining to what extent the variable is not correlated with 'g'. This follows necessarily from the fact that, when the criterion is satisfied, each variable is divisible into two parts, the general and the specific. Now, the extent to which any variable is not correlated with 'g' is found, after its correlation with 'g' has been computed, by substituting the latter value in the formula for the 'coefficient of alienation' (16, p. 173). For this purpose we have available the data of Table V, from which the following specific correlations are obtained:

TABLE VIII
CORRELATIONS OF NINE VARIABLES WITH THEIR SPECIFIC FACTORS

Vocabulary	.4430
Analogies	.5781
Sentence Completion	.7939
Disarranged Sentences	.9233
Arithmetic Reasoning	.6421
No. Series Completion	.8290
Equation Relations	.8872
Multiplication	.8736
Verbal Battery	.3667
Numerical Battery	.5587

Most of these correlations are sufficiently high to be quite impressive. Despite the fact that the verbal battery has a correlation of .93 with V, there is nevertheless a high degree of specificity in two of these tests, and a substantial degree of specificity in the others. This is even more noticeable in the numerical battery. In three of these tests, the specific ingredient overshadows the general factor markedly, so far, at least, as the correlations are concerned.

4. *V and N Scores*

Knowing the correlation of each of the tests with its central factor, we may now take the next step toward the determination of the correlation of V with N. It is necessary to find a V score and an N score for each of the subjects of the experiment. Since the obtained scores are weighted unequally, they cannot be combined into a composite score directly. It is first required that the weight of each separate score be reduced to 1.00. Such weighted scores are known as 'reduced' scores. Woodworth's method of weighting was used in computing reduced scores (9, p. 283). It consists simply in finding the difference between each obtained score and the average of the

distribution, and dividing this remainder by the standard deviation of the distribution. Having done this, we were then ready for the following step. We cannot combine reduced scores into a composite score, for this would not yield V and N scores, since we know that the test scores contain both general or central factors and specific factors. We need to separate out the central factor from each of the reduced scores. The resulting scores may then be combined into a composite which will be the score required.

With the correlation data available in Table V, two regression equations were written, one for V, the other for N. This necessitated the calculation of third order partial correlations, the partial standard deviations and regression coefficients. It will be noted that it is the reduced scores from which V and N are to be separated out. Since these reduced scores were obtained by dividing by the obtained standard deviations, they are all of equal weight, and it is therefore unnecessary to use the obtained standard deviations, in computing the partial standard deviations. All standard deviations, then, have a weight of 1.00, and the values obtained from the regression equations can then be applied directly to the reduced scores without overweighting or underweighting. Following this method, then, the regression equations are:

$$V = .5786X_2 + .2912X_3 + .1341X_4 + .0723X_5$$

in which X_2 is Vocabulary, X_3 is Analogies, X_4 is Sentence Completion, X_5 is Disarranged Sentences.

$$N = .3672X_2 + .2858X_3 + .2112X_4 + .1759X_5$$

in which X_2 is Arithmetic Reasoning, X_3 is Number Series Completion, X_4 is Equation Relations, X_5 is Multiplication.

The verbal tests contribute to V in the proportions given above. These proportions are, roughly, in the order given, 8 to 4 to 2 to 1. For the numerical tests, the relative contributions to N are, in the order given, roughly 2.0 to 1.6 to 1.2 to 1.0. In this scheme, the X_5 variable is weighted 1.

If the reduced scores referred to above are multiplied by the weights given in these regression equations, and the resultant scores combined, we shall then have for each subject a V score and an N score, in terms of equated units. This done, we need simply to compute the correlation between the V and N scores, the result being the correlation between V and

N, which is what we set out to find. The necessary computation having been made, it is found that the correlation between the central verbal factor (V) and the central numerical factor (N), is .2625, P. E. .0254. Since it has been shown that the Vocabulary Test is a highly efficient measure of V, it must follow that its correlation with N should approximate the correlation of V with N. This correlation was calculated, and it was found to be .2685. Comparison of this coefficient with that for V and N shows that the difference is very slight.

It will surely be agreed that there is little correspondence between the central verbal factor and the central numerical factor. Yet it has been demonstrated that the verbal tests, taken as a team, satisfy the tetrad difference criterion. This is true, also, for the numerical tests. The lack of correspondence between V & N cannot be due to specific factors, for these have been eliminated. It must be concluded, then, that whatever the nature of V and N, *they are assuredly not the same*. Indeed, it is not unreasonable to ascribe at least part of the correlation of $.2625 \pm$ P. E. .0254 to the fact that in *both* sets the subjects worked with paper and pencil, in the same room, under much the same conditions, *language being used throughout*. This similarity of conditions and partial similarity of content should of itself yield some degree of correlation, i.e., a small general factor, quite independently of similarity of mental function. According to Spearman, the central or general factor is the same throughout all cognitive functions, varying only in the degree to which particular functions draw upon it. It appears, then, that the outcome of this experiment is in conflict with Spearman's interpretation, or else it reduces our 'g' to a relatively small factor, probably verbal, for the most part.

There may be some objection to the foregoing conclusion on the ground that we ought to have calculated the full number of tetrad differences for the entire group of eight tests. It will be recalled that only six tetrad differences were calculated, whereas the full correlation table requires the calculation of 210 tetrad differences. Spearman requires that each and every one of these 210 tetrad differences equal zero before it can be said that the functions tested are divisible into two factors. There is some force in this objection, for it may be that the low correspondence between V and N is due to the presence of group factors, which Spearman interprets as the result of the

overlapping of specific elements. The criterion for the presence of such group factors is the failure of the tetrad difference criterion. This point, then, requires investigation.

E. Group Factors

If the verbal tests contain a group factor which is not present in the numerical tests, the application of Kelley's technique ought to reveal it (15, Prop. 16, p. 69). This proposition reads as follows:

"If the inter-correlations between four variables are such that $t_{1234} = t_{1243}$ and $t_{1342} = 0$, they could conceivably have arisen from four variables x_1, x_2, x_3, x_4 through which was a general factor plus, in addition thereto, a second factor common to x_1 and x_2 or a second factor common to x_3 and x_4 ." We shall apply this test to a table of intercorrelations involving Vocabulary, Analogies, Arithmetic and Number Series Completion. These are chosen because they are our best tests, in the sense that they have the highest correlations with their central factors. Let us build up the following theoretical scheme, after Kelley: If α = a general factor common to all four variables, β = an additional factor not shared by all of the variables, e = a factor or factors specific to each variable, and C and K = constants which do not change from one variable to another, we may then write:

- | | |
|--------------------------|--------------------------------------|
| 1. Vocabulary | $= C_1 \alpha_1 + K_1 \beta_1 + e_1$ |
| 2. Analogies | $= C_2 \alpha_2 + K_2 \beta_2 + e_2$ |
| 3. Arithmetic Reasoning | $= C_3 \alpha_3 \quad \quad + e_3$ |
| 4. No. Series Completion | $= C_4 \alpha_4 \quad \quad + e_4$ |

This gives us a situation in which we have a general factor present in all of the variables, and an additional factor present in only two of the tests, Vocabulary and Analogies. Correlations are obtained by multiplying the variables by one another. The e 's are, by definition, specific and therefore uncorrelated, and may therefore be disregarded. The c 's and K 's are constants, and therefore do not influence the correlations. By multiplication, then, we get the following correlations:*

* Since β does not correlate with anything in the number tests this term disappears from all the correlations except the first.

$$\begin{aligned}
r_{12} : \text{Vocabulary and Analogies} &= a_1 a_2 + \beta_1 \beta_2 \\
r_{13} : \text{Vocabulary and Arithmetic Reasoning} &= a_1 a_3 \\
r_{14} : \text{Vocabulary and No. Series Completion} &= a_1 a_4 \\
r_{23} : \text{Analogies and Arithmetic Reasoning} &= a_2 a_3 \\
r_{24} : \text{Analogies and No. Series Completion} &= a_2 a_4 \\
r_{34} : \text{Arithmetic and No. Series Completion} &= a_3 a_4
\end{aligned}$$

Applying the tetrad difference method, we obtain the following:

$$\begin{aligned}
t_{1234} &= a_3 a_4 \beta_1 \beta_2 \\
t_{1243} &= a_3 a_4 \beta_1 \beta_2 \\
t_{1342} &= 0
\end{aligned}$$

That is to say, the first two tetrads equal each other, while the third equals zero. This scheme has been applied to all the possible combinations of the four variables. In each of these combinations we have assumed the existence of the additional factor in two of the variables, but not in the other two. Finally we assumed an additional factor in the verbal tests, and still another additional factor present in the numerical tests, but absent from the verbal. These combinations have been worked out, in the manner illustrated above, the tetrads which ought to result being presented in Table IX.

TABLE IX
THEORETICAL TETRAD DIFFERENCES WHICH SHOULD RESULT WHEN CERTAIN CONDITIONS CONCERNING GROUP FACTORS ARE SATISFIED

<i>Assuming a Group Factor in:</i>	<i>Tetrads Should be:</i>
1. Vocabulary and Analogies	$t_{1234} = a_3 a_4 \beta_1 \beta_2$ $t_{1243} = a_3 a_4 \beta_1 \beta_2$ $t_{1342} = 0$
2. Arithmetic and No. Series Completion	$t_{1234} = a_1 a_2 \beta_3 \beta_4$ $t_{1243} = a_1 a_2 \beta_3 \beta_4$ $t_{1342} = 0$
3. Analogies and Arithmetic	$t_{1234} = 0$ $t_{1243} = a_1 a_4 \beta_2 \beta_3$ $t_{1342} = a_1 a_4 \beta_2 \beta_3$
4. Vocabulary and Arithmetic	$t_{1234} = a_3 a_4 \beta_1 \beta_3$ $t_{1243} = 0$ $t_{1342} = a_2 a_4 \beta_1 \beta_3$
5. Vocabulary and Completion	$t_{1234} = 0$ $t_{1243} = a_2 a_3 \beta_1 \beta_4$ $t_{1342} = a_2 a_3 \beta_1 \beta_4$

6. Analogies and No. Series Completion	$t_{1234} = \alpha_1\alpha_3\beta_2\beta_4$
	$t_{1243} = 0$
	$t_{1342} = \alpha_1\alpha_3\beta_2\beta_4$
7. Assuming a group factor (β) in Vocabulary and Analogies and an additional group factor (γ) in Arithmetic and No. Series Completion	$t_{1234} = \alpha_2\alpha_4\beta_1\beta_2 + \alpha_1\alpha_2\beta_3\beta_4 + \beta_1\beta_2\beta_3\beta_4$
	$t_{1243} = \alpha_2\alpha_4\beta_1\beta_2 + \alpha_1\alpha_2\beta_3\beta_4 + \beta_1\beta_2\beta_3\beta_4$
	$t_{1342} = 0$

Having worked out the tetrad differences which are required by theory in the seven situations presented in the foregoing table, it is in order to calculate the actual tetrad differences derivable from the four variables under consideration. These tetrad differences are given below:

$$\begin{aligned} t_{1234} &= .3120 \\ t_{1243} &= .3132 \\ t_{1342} &= .0012 \end{aligned}$$

It is clear that the first two tetrad differences may be taken as equal to each other, and the third as equal to zero. Within the limits of the probable errors, the situation conforms to Kelley's sixteenth proposition. If we think of the four variables as containing a general factor running through all, and an additional factor present in the two verbal tests, but not present in the numerical tests, we find that the obtained tetrad differences agree with the theoretical prediction. If we assume that the additional factor is present in the numerical tests, but not in the verbal, the facts also fit the theory. Lastly, if we assume an additional factor present in each of the two types of tests, there is still agreement with the theory. Of particular significance is the fact that no group factor cuts across the two forms of tests. There is no single group factor present in both the verbal and the numerical tests. This does not mean that we can be certain of two group factors, though two may exist. It is practically certain, however, that at least *one* group factor exists, either in the verbal or in the numerical tests. Complete certainty is not yet possible, since the converse of Kelley's proposition has not been proved. But the probabilities are over-whelmingly in favor of the presence of at least one group factor, confined to either the verbal or the numerical tests.

These conclusions are in agreement with the low correlation between the verbal and numerical abilities, as expressed in the

coefficient of .2625. Our original conclusions appear to be substantiated. Whatever the nature of V and N, the two are not the same. If the group factor is in the verbal tests, it cannot be due to the overlapping of specific factors, for the entire battery of verbal tests satisfies the tetrad difference criterion, i.e., all of the tetrad differences are zero, within the limits of probable error. The same observation applies equally if the group factor be located in the numerical tests. It still holds if there be two group factors, one in each battery of tests. The tetrad differences are not zero when the verbal and numerical tests are taken together, and this, according to Spearman, proves the existence of group factors. Since each battery, taken alone, satisfies the criterion, and since the criterion fails only when we cut across the boundary between the two batteries, it follows that the group factor should also cut across the boundary. But it has just been shown that this is not the case. In Table IX we have shown the statistical results which should be found if the group factor cuts across the two types of tests. These results, however, are not found. The observed facts will agree only with the theoretical predictions which are based on the assumption of a verbal factor, or a numerical factor, or both. There is no agreement with the assumption of a verbal-numerical factor, over and above 'g'.

3. Relation of V and N to Other Abilities

It has been shown that V and N are distinctive in the sense that their correspondence is quite low. There arises the further question as to whether they are also distinguishable from other abilities. It is possible that they may be nothing more than measures of what may be called 'general ability', and not distinctive traits. The answer to this question may be found by testing our group in other abilities, and then correlating these abilities with V and N. One such test has been made, in the field of memory.

Several weeks prior to the major experiment, a paired associates test in memory was given to the same group. The test was given in laboratory, in groups of 18 subjects. Absences reduced the size of the group from 210 to 187. The material was paired as follows:

pictures and pictures
words and words
words and nonsense syllables
geometrical forms and digits'

Two forms of the test were given visually, twenty pairs being contained in each of the forms. In each form, there were five pairs of each type as described above, making a total of ten pairs of each kind, and a total of forty pairs in all. The pictures were taken from periodicals. They were all in black and white, all of the same size, and all presented on cards measuring nine inches by four inches. They were pictures of common objects, such as a piano, a pencil, a watch, a dog, a boy, a bed, a shoe, and the like. The words were common four-letter words, without obvious associations. A few examples are: foot, past; four, bade; foal, high; trim, lark; fish, show. Some examples of the words and nonsense syllables are given: ring, bef; rake, dro; duck, orp; safe, ilt; the rest were of the same form. The geometrical patterns and digits were paired after the following manner; triangle, 56; oval, 27; diamond, 12, etc. All pairs were affixed to cards of the size mentioned for the pictures. The words, nonsense syllables and digits were all of the same size, printed on white cards, in black ink. The geometrical forms were approximately equivalent in area, inked in black and affixed to white cards. The exhibits were presented in irregular order, each being shown during four beats of a metronome set at 60.

The test was for immediate retention. The cards were presented a second time, the right side being in each case covered, only the left side being visible to the subjects. The subjects were required to state, in writing, what had been shown on that side of the card which was now covered. Both forms of the test were given during the same class period. Scoring was in terms of the number of items remembered. The results yielded the following table of correlations:*

TABLE X
CORRELATIONS BETWEEN VERBAL, NUMBER AND MEMORY TESTS

<i>Memory</i>	
Memory	.750 (self-correlation)
V	.114
N	.171
Synonyms	.070
Analogies	.115
Sentence Completion	.197
Disarranged Sentences	.067
Arithmetic Reasoning	.133
Number Series Completion	.164
Equation Relations	.086
Age	.077

* The writer is indebted to Mr. Murray Poliwinchik for the calculation of these coefficients.

The correlation between memory and memory (.750) is the reliability coefficient, found by correlating the two forms against each other, and applying the Spearman-Brown Formula. The other correlations were found by using, for each subject, his total score for each form of the memory test, and then combining these into a single score for the entire memory test, whose reliability has been given above. The reliability coefficient is not as high as might be desired, but it is, nevertheless, of reasonable magnitude. A practice effect is indicated by increased scores for the second form of the test. It is felt that the reliability of the test might be higher if preliminary practice had been given.

In no case is the correlation between memory and the other variables as high as .20. This points very clearly to the relative independence of the verbal and numerical abilities from memory, as measured by the test employed. This is the more striking, in view of the fact that the tests for retention involved words and numbers. Even the pictures and geometrical forms were probably dealt with verbally during the course of the experiment. In any case, the responses in the test for retention were written, and always in the form of a word, nonsense syllable or number. Even this community of form of content did not suffice to give appreciable correlations. One may, perhaps, be pardoned a guess that whatever correlation is present is due, in part at least, to similarity of content. It is conceivable that a test for retention which excluded words, syllables and digits might have yielded even a lower correlation with V and N.

4. Relations of Verbal and Numerical Abilities to Scholastic Records.

The writer believes that, so far as this study is concerned, the relative independence of V and N has been established. It was thought desirable, nevertheless, to compare the results with the scholastic records of the subjects. It is reasonable to expect some correspondence between performance in the tests and the work done by the students in their college courses. The validity of the conclusions already offered does not depend upon the existence of such correspondence, for it is well known that scholastic grades in college are by no means a consistently accurate criterion of the students' abilities. Many factors other than scholastic ability play a part in the determination

of grades. On the other hand, it is felt that the tests used in this study have yielded a fair gauge of ability in the functions tested, principally because of the novelty of the situation and the fact that the work was difficult enough to be challenging to the subjects. A close correspondence between performance in the tests and scholastic records is not to be expected. There should, nevertheless, be a trend, however slight, in the direction of correspondence. At any rate, it was thought that the comparison might be interesting, and it was, therefore, made.*

Scholastic records for the group used in the retention test were gathered from the files of the Registrar of the College. The record of one man was missing, so that the group was reduced in number from 187 to 186. The comparison was made in three ways:

1. The subjects' grades in literary and in science courses were collected. Literary courses were defined as those which involved direct study of language and literature. All courses in English, French, German, Spanish, Italian, Latin and Greek were included in this group. For the science group, grades were taken for all courses in Mathematics, Physics, Chemistry, Astronomy, Engineering and Biology. Letter grades were converted into numerical values by the following arbitrary weighting:

Grade A = 4 points

Grade B = 3 points

Grade C = 2 points

Grade D = 1 point.

The number of credits received by the subject in a given course was multiplied by the value given above. This yields a weighted value in points for each course. These weighted values were summed separately for the literary courses and for the science courses, giving, for each man, a total score for literature and for science. Each man's literature score was then divided by his total number of semester credits in literary courses, the result being an average value for the literary courses. The same procedure was followed with respect to the science courses. The subjects were then divided into three groups, based upon their average grade values. Group A stands high-

* The writer acknowledges his indebtedness to Mr. Murray Poliwinchik, who gathered the scholastic records from the files of the Registrar of the College. Mr. Poliwinchik also made most of the calculations recorded in the tables following.

est, Group B is the middle group, and Group C is the lower group. There are, of course, three groups under the heading of literature, and three under the heading of science. For each of these six groups, the average V and N scores were computed. It will be recalled that V and N scores were expressed in terms of sigma values. This means that they are fractional values, some positive and some negative. For the sake of simplicity of presentation, these sigma scores were converted in the manner described by Hull (9). Assuming a normal distribution with a mean of 50, a range of 0 to 100, and a standard deviation of 14, sigma scores may be converted readily into values in such a distribution. This, accordingly, was done, and the verbal and numerical scores given below are expressed in terms of such values.

TABLE XI-a

COMPARISON OF V AND N SCORES WITH ABILITY IN LITERARY AND SCIENCE COURSES

* Group A = highest grades. Group B = middle grades. Group C = lowest grades.

LITERARY COURSES				SCIENCE COURSES			
Group	N**	Verbal	Numerical	Group	N**	Verbal	Numerical
A	57	56.5	51.3	A	53	53.1	54.1
B	90	49.2	48.3	B	85	50.2	50.4
C	39	43.8	50.6	C	48	49.6	46.7

Analyzing the data for the literary courses, it is found that Group A, which is of superior ability in literary courses, has a higher average V score than does either of the other groups. The V score for Group A is also higher than the N score for the same group. The middle group, B, has a higher V score than does Group C, but its V score is only slightly higher than is its N score. This is not surprising, for it is not to be expected that a group which is mediocre in literary courses should be better in these than in science courses. Group C has the lowest V score, which was to be expected. Since this group is inferior in literary courses, one need not be surprised to find its N score higher than its V score; and this is precisely what is found. These results are clearly indicative of a trend towards correspondence between V and ability in literary courses.

* A, B, C are not grades. These letters merely designate the high, low and middle groups.

** N here refers to population.

Turning to the science group, a similar trend is noted. Group A has a somewhat higher score for N than for V. It is also the best group in N. Group B has a higher N score than does Group C, but, as before, there is practically no difference between V and N for this group. Group C has the lowest N score, and its V is higher than its N. The trend of results for the science courses parallels that for the literary courses quite closely. Comparing the literary with the science group, the expected results are found. Group A (literary) has a higher V score than does Group A (science), but a lower N score. Group C (literary) being inferior in literary courses, has a lower V score than does Group C (science), but a higher N score. Since the B groups are mediocre, there is no reason to expect any advantage in either direction. The B (science) group does have a higher V score than does the B (literary) group, but it has, also, a higher N score. Though these differences are small, they nevertheless constitute a departure from the otherwise uniform trend in the expected direction. This deflection, however, is not adverse to the general drift. It is simply outside the current. On the whole, the comparison points to some agreement between performance in the tests and performance in college work. This agreement, however, is not close.

2. Comparison with scholastic ability was made in another way. The group of 186 men was divided into two classes, the literary and the scientific. The literary group includes those men who have a greater number of semester credits in literary courses than in science courses. The scientific group includes those who have a preponderance of credits in science courses. Each of these groups was sub-divided into three parts, A, B and C, based upon the number of credits earned. It was thought that mere exposure to literary courses or to science courses ought to be reflected in the V and N scores. The C group includes those who had earned from 11 to 24 credits; the B group, from 25 to 38 credits; the A group, from 39 to 52 credits. The division was made in this manner for both the literary and the scientific men.

The average V and N scores were computed for each of the six sub-groups, with results as shown in Table XI-b.

The trend is fairly marked. The A and B literary groups have noticeably higher scores for V than for N. The A and B science men have distinctly higher scores for N than for V.

TABLE XI-b
COMPARISON OF V AND N SCORES WITH TRAINING IN LITERARY AND
SCIENCE COURSES

A, B, C, are defined as in Table XI-a.

LITERARY COURSES				SCIENCE COURSES			
Group	N	Verbal	Numerical	Group	N	Verbal	Numerical
A	26	54.4	48.4	A	33	48.0	52.8
B	40	54.5	48.6	B	39	48.2	52.2
C	30	46.2	46.5	C	18	52.3	48.8

The C literary men are very slightly superior in N than in V, while the C science men have higher V scores than N scores. These results agree very well with the hypothesis that mere exposure to literary courses ought to be reflected in higher V scores than N scores, regardless of grades. Similarly, a preponderance of work in science courses results in better N scores than V scores regardless of grades. The outcome is the same when the literary men are compared with the science men. The V scores for the former are higher than those for the latter, at least for the A and B groups. Conversely, the N score for the science men is higher than that for the literary men, Group C excepted. Though the C science men had more work in science than in literature they nevertheless appear to have had too little of science to develop superiority in N. It seems probable that training in other subjects, most of them involving verbal elements, had more effect than their relatively meagre training in scientific work.

3. Comparison of V and N with scholastic records was made in still a third way. The entire group was divided into two classes as before, literary and scientific men, based upon the field in which a preponderance of work was done. Instead of dividing each of these two classes into three sub-groups, based upon the number of credits earned, the division was made on the basis of average grades. Thus, the literary men were divided into three groups, A, B and C, the A group including the literary men who had the highest grade in literature, the C group including those literary men who had the lowest grades, the B group being intermediate. A similar division was made for the science men. It was thought that, of those men who had done a preponderance of work in literature, those who had earned the best grades ought to have the best V scores, while the best science men ought to have the best N scores. The results are shown in Table XI-c.

It is found that for the literary men, the V scores progress

TABLE XI-c
COMPARISON OF V AND N SCORES FOR LITERARY AND SCIENCE MEN
WITH THEIR GRADES IN THEIR RESPECTIVE FIELDS

<i>LITERARY MEN</i>				<i>SCIENCE MEN</i>			
<i>Group</i>	<i>N</i>	<i>Verbal</i>	<i>Numerical</i>	<i>Group</i>	<i>N</i>	<i>Verbal</i>	<i>Numerical</i>
A	30	53.2	52.6	A	33	50.2	53.6
B	41	47.3	47.3	B	43	48.5	51.7
C	25	44.4	47.8	C	14	46.7	46.9

downward as we proceed from Group A to Group C. A similar result is found for the science men. Their N scores become lower as we proceed from Group A to Group C. It is curious, however, that in the literary group, differences between the V and N scores are practically non-existent, excepting for the C men, and here the difference is, as before, in the reverse direction. On the whole, it appears justifiable to conclude that V and N scores are, in slight degree, in agreement with the quality of work done in corresponding courses in college. It would be premature to say that there is a close interdependence. It will be noticed that division of the experimental group into sub-groups resulted in very small populations, from which it would be rash to draw any far-reaching conclusions. There is, undoubtedly, a drift in the direction of agreement, but the outcome is, at best, only suggestive.

It would, indeed, be somewhat surprising if the agreement were much closer than has been found in this study. After all, there are many courses, other than the strictly literary, which are primarily, if not wholly, verbal in content. On the other hand, all courses in science, even mathematics, involve considerable verbal elements.* Where there is so much overlapping, it would be astonishing if clear-cut distinctions could be made. It would seem worth-while to submit the question to carefully controlled experiment. It is believed that the evidence here adduced is sufficient to warrant the tentative hypothesis that V and N are, in part, determined by training in verbal and numerical tasks. The most convincing evidence is contained in Table XI-b, from which, upon comparison with the other tables, it appears that there is closer correspondence of V and N with amount of work done, rather than quality of work. This last fact is not astonishing, for the unreliability of college grades is notorious.

* It is suggested that science men require reading comprehension, and this would tend to obscure the difference between verbal ability and numerical ability. On the other hand, numerical ability is not requisite for success in V.

V

SUMMARY

1. The problem was concerned with the inter-relationship existing between verbal and numerical abilities; the relation of these to other abilities; and the determination of the best instruments for measuring verbal and numerical abilities.

2. The subjects were 210 men representing a minimum of diversity with respect to education, ancestry, economic and social status, and age.

3. Nine tests were given, five verbal and four numerical, group method used.

4. General factors for verbal ability (V) and for numerical ability (N) were found.

5. Verbal and numerical ability were found to have little in common, the correlation between V and N being .2625.

6. It was established that either V or N contains something which is not present in the other.

7. The paired associates test for immediate retention was given and it was found that V and N, besides being differentiated from each other, are also distinguishable from retention. V and N are not, therefore, merely measures of general ability.

8. There is a noticeable but not marked correspondence between verbal ability and performance in literary courses in college. The same condition holds with respect to numerical ability and the performance in science courses, including mathematics.

9. The best single test of verbal ability is a vocabulary test. The best single test of numerical ability is the Arithmetic Test. For each ability the entire battery of tests is more efficient than any single test.

VI. BIBLIOGRAPHY

1. Anderson, R. C. 1925. A critical analysis of test-scoring methods. Archives of Psychology No. 80.
2. Burt, C. 1916. Distribution and relation of educational abilities. London County Council.
3. Burt, C. 1922. Mental and scholastic tests. London. King.
4. Bonser, F. G. 1910. The reasoning ability of children of the 4th, 5th, and 6th grades. T. C. Cont. to Education No. 37.
5. Bronner, A. F. 1917. The psychology of special abilities and disabilities. Boston, Little, Brown & Co.
6. Davey, C. M. 1926. Comparison of group verbal and pictorial tests of intelligence. Brit. J. Psych. 17, 27-48.
7. Dodd, S. C. 1928. The theory of factors. II. Psych. Rev. 35, No. 4, 268-279.
8. Dodd, S. C. 1928. The theory of factors. I. Psych. Rev. 35, No. 3, 211-234.
9. Garrett, H. E. 1926. Statistics in psychology and education. New York. Longmans, Green & Co.
10. Gates, A. I. 1922. Correlations of achievement in school subjects with intelligence tests and other variables. J. Ed. Psych. 13, 129-139.
11. Gates, A. I. 1922. Psychology of reading and spelling. T. C. Cont. to Educ. No. 129.
12. Gates, A. I. 1928. The improvement of reading. New York. Macmillan.
13. Hollingworth, L. S. and Winford. 1918. The psychology of special disability in spelling. T. C. Cont. to Educ. No. 88.
14. Holzinger, K. J. 1929. On tetrad differences with overlapping variables. J. Ed. Psych. 20, No. 2, 91-97.
15. Kelley, T. L. 1928. Crossroads in the mind of man. Stanford University, Stanford University Press.
16. Kelley, T. L. 1924. Statistical method. New York. Macmillan.
17. Spearman, C. 1927. The abilities of man. New York. Macmillan.
18. Spearman, C. 1912-1913. Correlation of sums and differences. Brit. J. Psych. 5, 417-426.
19. Symonds, P. M. 1923. Special abilities in algebra. Teachers' College, Columbia University, New York.
20. Terman, L. M. 1916. The measurement of intelligence. New York. Houghton, Mifflin Co.
21. Thomson, G. H. 1927. The tetrad difference criterion. Brit. J. Psych. 17, 235-255.
22. Thorndike, E. L. 1927. The measurement of intelligence. New York. Teachers' College, Columbia University.
23. Thorndike, E. L. 1921. Intelligence and its measurement: A symposium. J. Ed. Psych. 12, 127-129.
24. Wilson, E. B. 1929. Comment on Professor Spearman's note. J. Ed. Psych. 20, No. 3, 217-223.

VITA

The author, Matthew Maximilian Rupprecht Schneck, was born May 10, 1898, in San Bernardino County, California. He received the degree of A.B., with highest distinction, from the University of Arizona in June, 1925. The degree of Master of Arts was conferred upon him by Columbia University in the City of New York in June, 1927. Previous publications are:

1. A study of the discriminative value of the Woodworth Personal Data Sheet. With H. E. Garrett. Journal of General Psychology, 1928, 1, 459-471.
2. Retention in animals: A critical survey. With C. J. Warden. Journal of Genetic Psychology, March, 1929.

A PRELIMINARY STUDY OF THE EFFECT OF TRAINING IN JUNIOR HIGH SCHOOL SHOP COURSES

BY
L. DEWEY ANDERSON

Submitted in partial fulfillment of the requirements for the degree
of Doctor of Philosophy in the Faculty of Philosophy,
Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, Editor

No. 109

NEW YORK
August, 1929

CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	5
Description of Groups.....	7
Measurement of Group Similarity.....	9
Grading Methods.....	10
Reliability of the Course Grades.....	12
II. RESULTS	
Practice Effects	14
The Effects of Two Courses on Each Other.....	18
Suggested Curriculum for Shop Training.....	19
III. RESULTS	
Similarities in the Courses.....	21
IV. RESULTS	
Relationship of the Amount of Practice to the Degree of Group Similarity.....	26
Conclusions	33
V. RESULTS	
The Effect of Longer Periods of Shop Training..	34
Conclusions	36
VI. SUMMARY	
Procedure	37
Results	37

SECTION I

INTRODUCTION

This investigation is a study of the comparative value of training in five junior high school courses, Mechanical Drawing, Sheet Metal, Woodwork, Electricity, and Printing. The purposes are:

1. To determine the differences in the quality of work performed in each course by two groups of students, one, the control group, with no training prior to that in the course in question (test course) and the other, the trained group, with previous training in one of the other courses (practice course). Or in other words, did the group with previous training obtain a higher or lower average grade than the group without training? For example, how did the average grade in Electricity, obtained by a group of students with ten weeks' previous training in Mechanical Drawing, compare with that of a group with no training?

2. To determine whether there is any relationship between the differences in the quality of work performed by the two groups in each test course and the number of factors characteristic of both the course and its practice course.

3. To study the differences in the average grades obtained by a group of students at the conclusion of ten, twenty, thirty, and forty weeks of training, in order to determine whether training in school courses of this type had a cumulative benefit.

The purposes relate to problems which have a direct bearing on two fields of research: (1) the proper formation of curricula from experimental data, and (2) the transfer of training. In regard to the first, the results of this research should be significant as indicating procedure and a tentative arrangement of the five courses in a curriculum. As far as the writer is aware, in industrial education there has been no attempt to secure quantitative evidence which would indicate what courses should be prerequisites of other courses. In academic education Rugg's "The Experimental Determination of Mental Discipline in School Studies"¹ approaches the problem of curricula making in terms of practice effects. In his research Rugg measured the extent to which training in one course, Descriptive Geometry, operated to increase efficiency in various other mental activities. While his investigation was not planned

¹ Rugg, H. O. The Experimental Determination of Mental Discipline in School Studies. Educ. Psych. Monographs, No. 17, 1916.

primarily to determine whether Descriptive Geometry should precede other courses, the results showed that a benefit occurred if such were the case. However, no attempt was made to determine whether this benefit was greater or less than similar benefits from other courses. The present investigation gives the effects of training in one course on the quality of performance in another course and vice versa for five courses. Some or all of these courses are included in junior high school curricula.

The author regards his results in the interpretation of transfer problems as exploratory and suggestive since (1) the practice of determining the equality of groups after their composition is inferior to the practice of selecting cases for the groups to be compared and (2) the number of cases in each group did not assure high reliability of results.

It is to be borne in mind that few studies in either of the two fields, transfer of training or the formation of curricula, have reported any attempt to secure group equality or to determine group similarity, and that very few have used good controls in a school situation without modifying it to such an extent that it approached a laboratory one. Also, in the attacks on the problem in a school situation, the effect of practice in one course on the quality of work done in another has been studied without the reverse being done; and, moreover, the differences in performance between groups have been based on subjective grading methods.

The value of this research rests on the fact that the effects of practice were studied in a new field on a group of supposedly closely allied courses. Furthermore, the courses were of such a nature that the problem of the relation of the amount of practice effect to the number of common factors was based, first on objective measures of school performance (grades), and secondly on a numerical count of objective factors common to the two courses, practice and test.

Previous investigations on the problem of transfer of training have shown, in general, that:

1. Transfer of training may evidence itself either as an improvement of the second activity or an impairment.
2. The amount of transfer effect assumes small proportions in comparison to the value of training in the influenced activity itself.
3. The more two situations are similar, the greater the amount of transfer, other things being equal.

These conclusions were verified in the present investigation in a relatively new field, i.e., shop training in a junior high school. There are several good summaries of the literature on transfer of training.^{2, 3}

Description of Groups

This investigation was conducted in the city of Minneapolis in a junior high school which offered shop work consisting of ten weeks' training of five one-hour periods per week in each of the five courses, Mechanical Drawing, Sheet Metal, Woodwork, Electricity, and Printing.⁴ The school program did not require that the student take these courses in a prescribed order. Also, the five shop laboratories did not accommodate an equal number of students. Because of these two facts, the method of rotation was by student, rather than by group, through the various courses.

All the boys entering the seventh grade in September, 1924, with a few exceptions caused by physical disability, were divided into five sub-groups numbering from twenty to forty cases. Each of these sub-groups was assigned to a course. At the expiration of the first period of ten weeks, the boys were again assembled and re-distributed into sub-groups. In the assignment of these new sub-groups to courses, care was taken that no student was returned to a course which he had attended previously. This same procedure was repeated at the end of each ten weeks' period, so that all the boys, with the exception of those who discontinued work before the expiration of the school year, took work in each of the five shop courses.

The students in these groups had:

1. An age range from twelve to seventeen years.
2. Similar academic training.
3. No previous formal training in shop courses.
4. No special interest in shop work as a group.

In studies of the value of practice it is necessary to have two groups for comparison, one group differing from the other in one phase of the training. The trained group has a par-

² Whipple, Guy M. The Transfer of Training. The Twenty-seventh Yearbook of the National Society for the Study of Education. Part II, pp. 179-209, 1928.

³ Rugg, H. O., *op. cit.*

⁴ Data secured in connection with the research on mechanical ability conducted by the Department of Psychology, University of Minnesota, under the auspices of the Committee of Human Migrations of the National Research Council.

ticular kind of training which the control group lacks. The effect of this training is determined by a comparison of the average grades of the two groups. The first five sub-groups in shop training during the first ten weeks' period were used as the control groups for the five courses. For example, the students who took Woodwork for the first ten weeks formed the control group for this course. The trained groups included those students taking the test course (Woodwork in the above example) subsequent to training in some other courses. If the effect of practice in the Electricity course was to be measured on the quality of performance in Woodwork, the average grade for the Woodwork control group was compared to the average grade of the group previously trained in Electricity.

The school program made it possible to secure data on all possible combinations of the five courses with one exception, Mechanical Drawing preceded by training in Printing. The nineteen practice-test course combinations were:

<i>Practice-Test Course Combinations</i>	<i>Abbreviations used in Text</i>
Electricity followed by Woodwork	E W
Printing followed by Woodwork	P W
Mechanical Drawing followed by Woodwork	M W
Sheet Metal followed by Woodwork	S W
Electricity followed by Sheet Metal	E S
Printing followed by Sheet Metal	P S
Mechanical Drawing followed by Sheet Metal	M S
Woodwork followed by Sheet Metal	W S
Printing followed by Electricity	P E
Mechanical Drawing followed by Electricity	M E
Sheet Metal followed by Electricity	S E
Woodwork followed by Electricity	W E
Electricity followed by Printing	E P
Woodwork followed by Printing	W P
Sheet Metal followed by Printing	S P
Mechanical Drawing followed by Printing	M P
Electricity followed by Mechanical Drawing	E M
Sheet Metal followed by Mechanical Drawing	S M
Woodwork followed by Mechanical Drawing	W M

To repeat, the average grade of the student in the last-named course (above) in each of these combinations was compared to the average grade in the same course made by the students of the control group.

In order to secure a greater number of cases in the trained groups, those individuals who had taken the practice course followed by the test course were included although they may have had some other course previous to the practice course. For example, one student took Electricity first and then Woodwork; another took Sheet Metal first, next Electricity, and then Woodwork. In both cases Woodwork followed Electricity. A check on this procedure was made, using only the cases with no training previous to the practice course. (Described on page 18.)

As stated previously, in most of the research on the effect of practice no attempt has been made to determine group similarity in ability or to control the composition of the groups by a selection of cases. It has been assumed that the two groups compared were equal in ability. In this investigation the degree of similarity of the control and trained groups for each practice-test course combination was measured to determine if the differences between the groups in average grades were caused by variations in practice or by an actual difference in the ability to turn out a good quality of work. The method used for determining the similarity of groups is discussed in the next division.

Measurement of Group Similarity

The decision on which characteristics the degree of group similarity should be measured was made with reference to the correlation coefficients in Table 1. These coefficients represent the degree of relationship found to exist between certain measures and shop efficiency grades in the group of subjects used in this study. The coefficient of chronological age with shop efficiency, $+.02$, indicated that within a group restricted as to school grade such as this one, age had no effect on the quality of work performed, and that, therefore, it was unnecessary to take age into consideration as a variable factor. The intelligence quotient, secured from scores in the Otis Self Administering Test, Form A, has a low relationship with shop efficiency. The coefficient of shop efficiency with the test battery composed of three tests, Paper Formboard, Spatial Relations, and the Minnesota Assembly⁵ was $+.64$, which became $+.77$, when corrected for attenuation. This coefficient was

⁵ Anderson, L. Dewey. The Minnesota Mechanical Ability Tests. The Personnel Journal, Vol. 6, No. 6, pp. 73-78, April 1928.

not high but its magnitude corresponded favorably to that given as justification for the use of standard intelligence tests in studies of a similar nature. Scores in this test battery were, therefore, used in determining the similarity of compared groups in mechanical ability. The difference in the average scores on this mechanical ability test battery made by the control and the trained groups was taken into consideration in judging the effects of practice in one course on the quality of work done in another course.

TABLE 1
CORRELATION COEFFICIENTS OF AGE AND TEST SCORES WITH GRADES IN SHOP EFFICIENCY*

<i>Shop Efficiency with Standing in</i>	<i>Correlation Coefficient</i>	<i>P.E.</i>
Age	+ .02	± .067
I.Q.	+ .21	± .065
Packing Blocks	+ .26	± .063
Card Sorting	+ .19	± .065
Spatial Relations	+ .53	± .051
Paper Formboard	+ .52	± .051
Stenquist Picture 1	+ .24	± .065
Stenquist Picture 2	+ .31	± .061
Minnesota Assembly	+ .55	± .047
Test Battery (Spatial Relations, Paper Form- board and Minnesota Assembly)	+ .64	± .039

* Measure of shop efficiency discussed on page 12.

Two comparisons were necessary, therefore, for each practice-test course combination: (1) a comparison of the average grades in shop work and (2) a comparison of the average scores in the mechanical ability test battery. A difference in the average grades, which indicated that one course had a favorable effect on another course, was not interpreted as such unless the average scores in the test battery were found to be equal or nearly equal. That is, when any difference in the average course grades was not accompanied by a similar tendency in the mechanical ability average scores, then the difference was interpreted as indicating that practice in one course had exerted an effect on the quality of work performed in a second course.

Grading Methods

Since ordinary school marks have a low reliability, their use in investigations of this kind should be avoided if possi-

ble.^{6, 7} Marks given by teachers, unless based on scores on objective tests, are likely to be biased because of personal likes and dislikes, or do not give a true indication of performance because other factors, such as earnestness, are taken into consideration. Since there is no fixed standard, the grades given by one teacher are not comparable to the grades given by another teacher, and the grades given by the same teacher on two different sections are not comparable.

This difficulty was overcome in the present investigation by measuring objectively the shop products made by each student in terms of deviations from standard products. The projects in each of the courses were:

Sheet Metal—foot scraper, tin box, biscuit cutter, cookie cutter, dustpan.

Woodwork—ruler, game board, necktie rack, final test project.

Mechanical Drawing—twelve project drawings made from isometric blue prints.

Printing—four “set-ups,” i.e., composition of sentences by use of type, leads, etc.

Electricity—One wiring project and three splices.

The students were required to make the above named projects from raw materials by the use of tools according to verbal instructions, actual demonstration of method, blue prints, or patterns. A fixed procedure was used in measuring each set of projects.⁸ For example, on the ruler made in the Woodwork course, measurements were made of the length, width, thickness, and squareness in order to determine the variations from blue print specifications. This type of measurement was possible for all the projects except those in Printing. These were scored on such characteristics as the transposition of letters, the use of wrong font, the evenness of margins, spelling, and justification. Since each measurement of each project was compared to a standard, the grades were in comparable terms.

This measurement procedure gave a series of scores for each boy in each project of each course. The needs of the present

⁶ Dearborn, W. F. School and University Grades. Bulletin of the University of Wisconsin, No. 368.

⁷ Banker, H. J. The Significance of Teachers' Marks. Journal of Educational Research, Vol. 16, pp. 159-71, Oct. 1927.

⁸ Anderson, L. Dewey. Measurement of Shop Products. Industrial Arts Magazine, Vol. XV, No. 8, pp. 263-267, 1926.

research required that these be combined in order to give a single grade index of the quality of work done on the projects in each course. It was necessary, therefore, to combine the measurements to give a project score, and then to combine the project scores to give a course grade. The progressive steps taken in treating the data were as follows:

1. Detailed measurements made on projects.
2. Project scores secured by combining the results of the detailed measurements.
3. Shop course grades secured by combining the grades on the projects within each course.

In the actual procedure it was necessary to place all scores in comparable units. When a measurement was completed the score for each individual was converted into sigma deviations from the average of the entire group. An average of these sigma deviations gave a numerical index of the quality of work done on each project.

The project scores for each individual were converted into sigma deviations from the average of the entire group and combined. The projects were weighted equally in the final course grade since the sigma deviation scores were in comparable terms. These course grades were used as the measures of proficiency.

The final grade in shop efficiency, referred to on page 10, and Table 1, were next secured by converting the course grades into sigma deviations from the average. Their total then gave for each student the objective grade which represented the quality of his work during the entire period of shop training.

In brief, the steps in this grading process were as follows: First, measurements on specific features of each project were made and converted into sigma deviations from the average of the whole group. These specific sigma deviation measures were averaged to give an individual's score for the project. The project scores were placed in terms of sigma deviations from the average of the whole group and averaged. Finally the course grades were converted into sigma deviations from the average of the whole group, and averaged to give a grade representative of the quality of work done by each individual in all the courses.

Reliability of the Course Grades

In the actual accomplishment of the above procedure, two grades were secured for each student in each course. The

first represented the sum of the odd numbered sigma deviation scores on the specific measurements; and the second, the sum of the even numbered sigma deviation scores as they appeared on the score sheets. Each grade, therefore, represented the average score on one-half of the measurements made on the projects made in each course. A correlation coefficient was computed between the two series of grades for each course. The reliability of the grading method was then obtained by treating this coefficient with the Brown Spearman Prophecy Formula, (Table 2). Four of the coefficients were large enough to permit group comparisons,⁹ but the fifth, for Electricity, was so low that the results for this course were questionable.

TABLE 2
THEORETICAL RELIABILITIES OF THE OBJECTIVE COURSE GRADES
(AVERAGE N—100)

<i>Course</i>	<i>Number of Measurements</i>	<i>Correlation Coefficient</i>	<i>P.E.</i>
Sheet Metal	64	+ .72	± .034
Woodwork	200	+ .76	± .029
Mechanical Drawing	300	+ .93	± .008
Printing	700	+ .69	± .034
Electricity	50	+ .35	± .059

In summary, by using the above procedure of subjecting the products of school work to actual measurement, the grades obtained were independent of personal judgment, were in terms of deviations from the average of the entire group of cases, were based on objective performance and had known reliabilities.

⁹ Kelley, T. Interpretation of Educational Measurements. World Book Company, pp. 210-11, 1927.

SECTION II

RESULTS—PRACTICE EFFECTS

The first purpose of this study was to determine the practice effects of each of five courses on one another. The following discussion is based on the differences in the average grades obtained in each course by two groups of students: one, the control group without training in other courses, and the other, the trained group, with training of a particular nature. For example, the average grade in Woodwork of all individuals who had Sheet Metal previously, was compared to the average grade in Woodwork of all students with no other shop training. While it is impossible to state definitely that a difference of this nature is an indication of practice effect, it is generally assumed that such is the case. The support of this assumption lies in the hypothesis that a variation in a situation is the cause of the change in performance. In the discussion of the results, the cause and effect terminology is used for greater objectivity and clarity. The term "positive effect" is used to indicate that the group with practice obtained a higher average grade than the group without practice. Conversely, "negative effect" refers to the situation in which the control group secured the higher grade, i.e., practice had an unfavorable effect.

A description has been given of the procedure used in determining whether a difference in course grades was due to (1) variations in training or (2) group dissimilarity in mechanical ability. In doing this the following situations were kept in mind.

1. A pronounced positive or negative difference in the average grades of the control and the trained groups when accompanied by none or a very slight difference in the average scores on the test battery was considered as indicative that the difference in achievement was due to previous training.

2. A pronounced difference in the course grades, either positive or negative, accompanied by a pronounced difference in the reverse direction in the mechanical ability averages was taken to indicate that the difference in average grades was the result of previous training.

3. A pronounced difference in the course grades, accompanied by a difference in the mechanical ability test averages in

the same direction was taken to indicate that the cause of the grade difference was indeterminate.

The data are given in Table 3. In this and succeeding tables a positive tendency (favorable practice effect), i.e., a difference that showed that the trained group scored higher on the average than the control group, is indicated by a plus sign. A minus sign indicates that the difference was in the reverse direction, i.e., that the control group had a higher average grade than the trained group.

An examination of the results in this table indicated in general that:

1. The direction of the differences in average grades was not always the same. Twelve of the differences were negative (the average grade of the control group higher than that of the trained group) and seven were positive.

2. The differences of the average scores on the mechanical ability test battery of the control and trained groups were not large. The random sampling method of selection gave similar groups.

3. Of the twelve negative tendencies in practice effect, four were accompanied by a lower mechanical ability average. Of the seven positive practice effect tendencies, two were accompanied by a higher mechanical ability score.

Since the indices showing the significance of the differences of the average grades were not high, the results could be interpreted as suggestive only. With this reservation in mind, the following specific tentative conclusions were made:

1. When Woodwork was the practice course, the trained group obtained a lower average grade than the control group in Printing, about the same in Sheet Metal, and higher average grades in Electricity and Mechanical Drawing. The differences in the mechanical ability averages did not influence these results.

2. When Sheet Metal was the practice course, the trained group obtained a lower average grade than the control group in Woodwork and Printing, and a higher average grade in Electricity and Mechanical Drawing. Again it was impossible to explain any of these results by a difference in the mechanical ability averages, except possibly that of Mechanical Drawing.

3. When Mechanical Drawing was the practice course, the trained group obtained a lower average grade in Woodwork, Sheet Metal, and Printing and a higher average grade in Electricity. The occurrence of these tendencies could not be explained by the differences in the mechanical ability averages.

TABLE 3

SHOWING (1) DIFFERENCES IN THE AVERAGE COURSE GRADES AND (2) DIFFERENCES IN THE MECHANICAL ABILITY TEST BATTERY AVERAGES FOR THE CONTROL AND TRAINED GROUPS FOR EACH PRACTICE-TEST COURSE COMBINATION

Practice-Test Course Combination	(1) Average Course Grades			Sigma of Diff.	(2) Average Test Battery Score		Diff.	Sigma of Diff.*
	Control No. Ave.	Trained No. Ave.	Diff. in Ave. Grade		Control Ave. Score	Trained Ave. Score		
WE	28 5.17	58 4.96	+ .21	1.18	2142	1899	-243	.26
WS	31 4.97	38 5.00	- .03	.19	2149	1995	-154	.17
WM	42 5.19	25 5.04	+ .15	.84	2178	2257	+ 78	.09
WP	25 5.04	19 6.15	-1.09	2.78	1735	2237	+ 502	.64
SW	25 4.81	52 5.12	- .31	1.89	2122	2141	+ 19	.02
SE	28 5.17	45 4.77	+ .40	2.22	2142	2027	-115	.12
SM	42 5.19	27 5.02	+ .17	.88	2178	2440	+ 263	.28
SP	25 5.04	40 5.34	- .30	1.03	1735	2363	+ 628	.77
MW	25 4.81	44 5.10	- .29	1.74	2122	2141	+ 19	.02
MS	31 4.97	42 5.10	- .13	.76	2149	2284	+ 134	.14
ME	28 5.17	48 4.72	+ .45	2.54	2142	2048	- 94	.10
MP	25 5.04	68 5.35	- .30	1.28	1735	2304	+ 569	.61
EW	25 4.81	23 5.36	- .55	3.09	2122	1855	-267	.29
ES	31 4.97	32 4.96	+ .01	.06	2149	2019	-130	.14
EM	42 5.19	28 4.91	+ .28	1.50	2178	2145	- 32	.03
EP	25 5.04	59 5.26	- .22	.92	1735	2253	+ 528	.57
PE	28 5.17	27 5.28	- .11	.51	2142	1775	-367	.39
PW	25 4.81	53 5.17	- .36	2.26	2122	1899	-224	.24
PS	31 4.97	54 5.09	- .12	.85	2149	2239	+ 90	.10

NOTE: Low numerical average grade indicates high actual grade. Signs reversed before differences in averages to show true tendency

* This refers to the sigma of the distribution of the entire group of subjects. This index is not to be confused with the sigma of the difference.

4. When Electricity was the practice course, the trained group obtained a higher average grade than the control group in Mechanical Drawing, nearly the same in Sheet Metal, and a lower average grade in Woodwork and Printing. The difference in the case of Woodwork was due probably to a similar difference in the mechanical ability test averages.

5. When Printing was the practice course, the trained group obtained a lower average grade than the control group in all courses, Sheet Metal, Electricity, and Woodwork. The results for the last two were due probably to corresponding tendencies in the differences of the mechanical ability averages. There were no data available for Mechanical Drawing.

The above conclusions were made from the standpoint of the same practice course and different test courses. When the data were studied from the standpoint of the practice effects of various courses on a single test course, (Table 3), the conclusions were:

1. The quality of performance in Printing was affected unfavorably by training in each of the four other courses.

2. The quality of performance in Woodwork was affected unfavorably by training in each of the four other courses.

3. The quality of performance in Sheet Metal was affected unfavorably by training in each of the four other courses.

4. The quality of performance in Mechanical Drawing was affected favorably by training in Sheet Metal, Woodwork, and Electricity. There were no data for Printing.

5. The quality of performance in Electricity was affected favorably by training in each of the four other courses.

The most significant practice effects were: the negative effects of Sheet Metal, Electricity and Mechanical Drawing on Woodwork, the negative effect of Woodwork on Printing, and the positive effects of Mechanical Drawing and Sheet Metal on Electricity.

In order to determine whether the use of all possible cases in the trained groups gave erroneous data, a comparison of the two kinds of trained groups was made and the results are shown in Table 4. Trained Group I, for each practice-test course combination, was composed of all possible cases, some of whom had had training in some other course before taking the practice course. Trained Group II was composed of those who had no other training but the practice course prior to the test course. The figures in Table 4 indicate that in these eight practice-test course combinations the use of the larger numbers of cases did not give erroneous indications. In no

TABLE 4
SHOWING DIFFERENCES IN THE AVERAGE COURSE GRADES BETWEEN THE CONTROL GROUP AND TRAINED GROUP I (ALL CASES) AND TRAINED GROUP II (SELECTED CASES) FOR EIGHT PRACTICE-TEST COURSE COMBINATIONS

Practice-Test Course Combination	Control No. Ave.		Trained				Diff. Between Control and Trained Groups	
			I No. Ave.	II		I	II	
				No. Ave.	No. Ave.			
W E	28	5.17	58	4.96	16	4.75	+.21	+.42
W S	31	4.97	38	5.00	16	4.96	— .03	+.01
W M	42	5.19	25	5.04	8	4.87	+.15	+.32
S W	25	4.81	52	5.12	12	4.97	— .31	— .16
S M	42	5.19	27	5.02	10	4.82	+.17	+.37
M E	28	5.17	48	4.72	26	4.92	+.45	+.25
E M	42	5.19	28	4.91	26	4.88	+.28	+.31
P W	25	4.81	53	5.17	14	4.92	— .36	— .11

NOTE: Low numerical average grade indicates high actual grade. Signs reversed before differences to show true tendencies.

case was the direction of the difference between the average grades of the two groups changed by using the smaller number of cases. Therefore the results secured from the larger number of cases are used throughout the investigation.

The Effects of Two Courses on Each Other

The problem here was to determine whether a course which had a practice effect on another course was affected in the same way by it. That is, did the same practice tendency occur regardless of which course was taken first? The data are given in Table 5.

Of the nine comparisons, six showed the same tendency in practice effect, irrespective of the order in which the courses were taken. That is, two courses affected each other in the same manner: either both the effects were positive or both negative. Three of the comparisons showed that the nature of the effect was different if the order was reversed; a positive tendency occurred under one arrangement and a negative tendency occurred under the reverse arrangement. One of these three was questioned because the same tendency was noted in the mechanical ability averages in at least one practice-test course combination. In other words, the differences in performance of the control and trained groups might have been due to group dissimilarity and not to a variation in practice. Two of the first six mentioned were doubtful because of this fact. In a great majority of cases, therefore, two courses have the same general practice effect on each other.

TABLE 5

SHOWING THE DIFFERENCES IN AVERAGE GRADES OF THE CONTROL AND TRAINED GROUPS (1) WHEN THE COURSES IN A PRACTICE-TEST COURSE COMBINATION ARE ARRANGED IN ONE ORDER AND (2) WHEN THE COURSES ARE ARRANGED IN THE OPPOSITE ORDER

<i>Courses</i>	<i>Diff. if First Named Course is Practice</i>	<i>Diff. if Second Named Course is Practice</i>
W E	+ .21	— .55*
W M	+ .15	— .29
W S	— .03	— .31
W P	— 1.09	— .36*
S E	+ .40	+ .03
S M	+ .17	— .13
S P	— .30	— .12
M E	+ .45	+ .28
E P	— .22	— .11*

NOTE: *Questionable as the differences in the mechanical ability averages were in the same direction.

Webb¹⁰ in an investigation of transfer of training in maze learning, using five human and nine animal subjects, made a somewhat similar conclusion. Six mazes were used for the animals and three for the humans. The mazes were paired and one group of subjects learned the mazes in one arrangement and another group learned the mazes in the reverse direction. His conclusion was that "the pair of mazes most similar yielded the smallest difference of saving when the direction of the transfer was reversed, while the largest difference in saving tended to obtain from the most dissimilar pair of mazes."

Suggested Curriculum for Shop Training

Since some of the courses have negative effects on others, this problem became one of securing the most benefit from training and at the same time minimizing as far as possible the unfavorable practice effects. If a course is beneficial to a second, it should precede the latter, or if its effect is negative in character, it should be taken later. The figures indicating the amount of practice effect (Table 6) are arranged in the summary below so that the most favorable tendency is given first for each pair of courses.

¹⁰ Webb, Louie Winfield. Transfer of Training and Retroaction. Psychological Monographs, Vol. XXIV, No. 3, 1917.

W S	— .03	S W	— .30
W M	.15	M W	— .29
W E	.21	E W	— .55
P W	— .36	W P	— 1.09
S M	.17	M S	— .13
S E	.40	E S	.03
P S	— .12	S P	— .30
M E	.45	E M	.28
P E	— .11	E P	— .22
M P	— .30	P M	no data

An analysis of these data indicated that Electricity should be preceded by four courses, Mechanical Drawing by three courses, Sheet Metal by two courses, and Printing and Woodwork by one course each. Electricity should have the last position in the curriculum because all the other courses had more favorable effects on it than it had on them. Mechanical Drawing likewise should be near the end, but before Electricity, because the effect of Mechanical Drawing on Electricity was more favorable than the reverse. Sheet Metal should have a position following Printing and Woodwork because it had more unfavorable effects on these courses than they did on it. Sheet Metal, as already shown, should precede Mechanical Drawing and Electricity. The effect of Printing on Woodwork was less unfavorable than the effect of Woodwork on Printing. The following order of courses was indicated by the analysis.

Course I	Printing
Course II	Woodwork
Course III	Sheet Metal
Course IV	Electricity
Course V	Mechanical Drawing

The position of Printing as the first course is doubtful since it affects all courses unfavorably. There is considerable argument in favor of placing this course last in the sequence since only one course would be affected unfavorably, while in the above order all the courses are affected unfavorably. In all probability Printing should not be included in a shop course curriculum. Emphasis should be laid on the fact that this analysis is based on indications of small practice effects either of a negative or a positive nature, and therefore should be regarded as tentative and suggestive as to method.

SECTION III

RESULTS—SIMILARITIES IN THE COURSES

The problem of determining whether or not the differences in practice effects were paralleled by the number of characteristics common to a test course and its practice course was next undertaken. Shop courses may be characterized in terms of specific movements performed, operations used, methods of instruction, use of models or patterns, materials worked on and product completed. The five shop courses were compared on these characteristics with the exception of that of specific movements. This was omitted, as its analysis would have entailed subjective judgments.

Operations

A list of the operations required in making the products in each course was compiled by the course teacher, with the aid of five other teachers. Measuring, boring holes, placing type in a case, folding wire, forming sheet metal, soldering, etc., were considered as operations. These were classed as operations rather than specific movements, as they are series of movements with a particular end in view. The lists were not the result of subjective judgments since only those operations which are overt and can be observed were included. If a further step had been taken in an attempt to state that measuring consists of the comprehension of a situation, attention to the task, eye-hand coordination, etc., then the realm of subjective decision would have been reached.

The operations lists for the courses were as follows:

Woodwork—chamfering wood, chiseling, drawing lines with a compass, drawing lines with a pencil, drilling holes, filing, gouging wood, hammering, measuring, planing wood, applying putty, sanding, sawing, inserting screws, spoke shaving, and varnishing. Total—16.

Mechanical Drawing—drawing lines with compass, drawing lines with pencil, lettering, locating drawings on page, measuring, and numbering. Total—6.

Printing—aligning type, selecting capital letters, justifying type, selecting slugs and leads, selecting type from cases, selecting right font, setting material on a base, squaring up composed material, and tying type form with string. Total—9.

Sheet Metal—beading metal, bending metal, cutting tin, cutting wire (pliers), drawing lines with compass, drawing lines with pencil, forming metal, filing, hammering, joining materials, measuring, punching metal, soldering, and inserting wire in tin fold. Total—14.

Electricity—connecting wire to posts, cutting wire, drilling holes, measuring, inserting screws, setting material on a base, skinning wire, splicing wire, and tracing electrical circuits. Total—9.

The number of operations common to each pair of courses were as follows:

Woodwork and Electricity	3
Woodwork and Sheet Metal	2
Woodwork and Mechanical Drawing	2
Woodwork and Printing	0
Printing and Mechanical Drawing	0
Printing and Sheet Metal	0
Printing and Electricity	1
Mechanical Drawing and Sheet Metal	2
Mechanical Drawing and Electricity	1
Sheet Metal and Electricity	2

Similar Tools

An analysis was made on the basis of the tools used in each course.

Woodwork—bit and brace, chisels, drawknife, file, claw hammer, knife, pencil, plane, ruler, saw, screw driver, spoke-shave, square and vise.

Mechanical Drawing—45 and 60 degree angles, compass, dividers, drawing board, eraser, pencil, ruler, and T-square.

Printing—brush, printing press, and sponge.

Sheet Metal—bending machine, compass, dividers, file, folding machine, machinist's hammer, hole punch, knife, pencil, pliers, rivet set, ruler, soldering iron, square, vise, and wiring machine.

Electricity—claw hammer, knife, pencil, pliers, ruler, screw driver, square, and vise.

The number of tools common to each pair of courses were:

Woodwork and Mechanical Drawing	3
Woodwork and Printing ..	0
Woodwork and Sheet Metal	7
Woodwork and Electricity	7
Mechanical Drawing and Printing	0
Mechanical Drawing and Sheet Metal	5
Mechanical Drawing and Electricity	3
Printing and Sheet Metal	0
Printing and Electricity .	0
Sheet Metal and Electricity	7

Similar Materials

The materials used in each course were:

Woodwork—glue, nails, screws, shellac, wood, and wood filler.

Mechanical Drawing—paper.

Printing—gasoline, ink, leads, paper, slugs, string, and type.

Sheet Metal—flux, sheet iron, paper, rivets, solder, and wire.

Electricity—insulators, nails, screws, light socket, light switches, wire, and wood.

The number of materials common to each pair of courses were:

Woodwork and Mechanical Drawing	0
Woodwork and Printing	0
Woodwork and Sheet Metal	0
Woodwork and Electricity	3
Mechanical Drawing and Printing	1
Mechanical Drawing and Sheet Metal	1
Mechanical Drawing and Electricity	0
Printing and Sheet Metal	1
Printing and Electricity	0
Sheet Metal and Electricity	1

Products

The products in Mechanical Drawing and Printing were of the one-plane type. In the other three courses, the products were three-dimensional. The Electricity products, however, were similar to the one-plane type since their construction involved placing objects on a base. Woodwork, Sheet Metal and Electricity were given one point each for similarity with one another. Printing, Electricity, and Mechanical Drawing were given one point each for similarity with one another.

Models and Patterns

The models or patterns, from which the students worked, varied from the actual product to blue print specifications. In the Mechanical Drawing course, the pattern was an isometric drawing, from which three one-plane drawings were made. The models for Sheet Metal and Woodwork were blue prints. For Electricity the models were actual products. The model for Printing was a composition on paper similar to the final product, yet different from a series of type locked in a form. This classification was not used, because only two courses had any point of similarity.

Methods of Instruction

In each course, before beginning a project, the students were assembled and the instructor demonstrated the product to be made, and showed the tools and how each was to be used in the various stages of making the product. Since the same procedure was followed in all the courses, this classification was not used.

Similar Tools and Similar Materials

The number of similar tools and the number of similar materials were added in order to secure the index of similarity for each practice-test course combination. Each tool and each material counted as one. The number of points of similarity for this and the remaining combination are given in Table 6.

Rank Order Total

The practice-test course combinations were ranked in order from the pair possessing the most points in common to the one possessing the least in the following classifications: similar operations, similar tools, similar materials and similar products. The ranks for each practice-test course combination were added, and the combinations reranked according to these final totals. This procedure was used because it gave each classification a weight of one in the final total.

TABLE 6
SHOWING THE GRADE DIFFERENCES BETWEEN THE CONTROL AND TRAINED GROUPS AND THE NUMBER OF COMMON FACTORS FOR THE PRACTICE-TEST COURSE COMBINATIONS ARRANGED ACCORDING TO THE SAME PRACTICE COURSE

<i>Practice-Test Course Combination</i>	<i>Diff. in Grades</i>	<i>Tools</i>	<i>Materials</i>	<i>Total¹</i>	<i>Operations</i>	<i>Total²</i>	<i>Product</i>	<i>Rank Order Total</i>
WE	+.21	6	3	9	3	12	1	10
WM	+.15	3	0	3	2	5	0	3
WS	-.03	7	0	7	4	11	1	8
WP	-1.09	0	0	0	0	0	0	1
SE	+.40	7	1	8	3	11	1	9
SM	+.17	5	1	6	3	9	0	7
SP	-.30	0	1	1	0	1	0	2
SW	-.30	7	0	7	4	11	1	8
ME	+.45	2	0	2	1	3	1	5
MS	-.13	5	1	6	3	9	0	7
MW	-.29	3	0	3	2	5	0	3
MP	-.30	0	1	1	0	1	1	4
EM	+.28	2	0	2	1	3	1	5
ES	+.03	7	1	8	3	11	1	9
EP	-.22	0	0	0	1	1	1	6
EW*	-.55	6	3	9	3	12	1	10
PE*	-.11	0	0	0	1	1	1	6
PS	-.12	0	1	1	0	1	0	2
PW*	-.36	0	0	0	0	0	0	1

NOTE: *This difference may be explained by a difference in the same direction in the mechanical ability test battery scores.

¹Tools and Materials.

²Tools, Materials and Operations.

SECTION IV

RESULTS—RELATIONSHIP OF THE AMOUNT OF PRACTICE TO THE DEGREE OF SHOP SIMILARITY

The differences in the average grades of the trained and control groups in the practice-test course combinations were compared with the number of common factors. Five methods were used:

1. The four practice-test course combinations with the same practice course were arranged in the order of the magnitude of the difference between the average grades of the control and trained groups. The number of factors common to the practice and the test courses were compared to these rankings to determine whether the order under one arrangement was similar to the order under the other.

2. The nineteen practice-test course combinations were arranged in order according to the magnitude of the differences between the average grades of the control and trained groups. Seven of the combinations showed positive practice effects and twelve, negative effects. Two procedures were followed:

- A. The average number of common factors in each classification was secured for those seven combinations which showed positive practice effects and for the twelve which showed negative practice effects. These averages were compared.

- B. The course combinations which showed positive practice effects were divided into two sections and the ones which showed negative effects were divided into three sections. The averages of the number of common factors in the sections were compared to the average practice effects to determine whether a decrease in the one was paralleled by a decrease in the other.

3. By using the same arrangement of practice-test course combinations as in (2), the percentages of the number of times that a combination having a higher practice effect standing than another, had more, the same, and less common factors were secured for the 171 possible comparisons.

4. The average amount of benefit (practice effect) each course received from the other four courses was secured, together with the average number of factors common to this course and its practice course. Two procedures were followed:

- A. A comparison was made to determine whether the courses receiving the greatest average benefit also had the greatest average number of common factors.

- B. The percentages were secured for the number of times a course received more practice benefit than another course and had more, the same, and less factors in common with the other courses.
5. The data were studied with the assumption that if two courses have many points of similarity, then, their practice effects on a third course should be approximately the same. The average amounts of fluctuation were secured for the courses most and least similar in the number of common factors.

Results. Method I

Table 6 contains the data. In the first column, the practice-test course combinations are given. In the second, are given the differences in the average grades of the control and the trained groups. The remaining columns contain the number of common factors in each classification, the totals for tools, materials and operations, and the Rank Order Total. The table is divided into five sections according to the practice courses. The four practice-test course combinations, with the same practice course, are arranged according to the amount of practice benefit each received.

If there were perfect correspondence between the amount of practice effect and the number of common factors, then, the figures should range from the greatest number to the least in each section. This was far from the case. The position of Printing, which showed in all practice-test course combinations the lowest amount of practice effect, was explained adequately by the non-occurrence of common factors. In general, Electricity, with the greatest benefit from practice in each practice-test course combination, also showed the greatest number of common factors.

The gradation of the differences of practice effect, when Woodwork was the practice course, was paralleled in the common factor classifications with the one exception of "tools" and the totals in which it was included. The correspondence was not marked so clearly when Sheet Metal was the practice course. One combination, Sheet Metal—Woodwork, was entirely out of position. For the other three courses, the correspondence was very good. When Mechanical Drawing was the practice course, the correspondence between practice effect and the number of common factors was not good because of the position of Electricity. When Electricity was the practice course, Mechanical Drawing received the greatest amount of

practice effect, but the number of common factors in any of the classifications was relatively low. The position of Woodwork, with the lowest ranking, might be explained in that the training group had much less mechanical ability than the control group. When Printing was the practice course, the amounts of practice effect paralleled fairly well the number of common factors. The rankings (Rank Order Total) showed the highest relationships to the amount of practice effect.

In general, the results indicated that the number of similar characteristics, measured in terms of tools, materials, operations, or a combination of these, had a positive relationship to the magnitude of practice effect. That is, the more factors two courses had in common, the more certain was the taking place of positive practice effect.

Results. Method II

Under the second method, the practice-test course combinations were arranged in order according to the magnitude of the differences in the average grades of the two groups, irrespective of the courses in the combinations. As already stated, seven of the combinations showed positive practice effects, and twelve, negative practice effects. The average number of common factors possessed by the practice and test courses were secured for those combinations which showed positive practice effect and for those which showed negative practice effect. The data are given in Table 7.

A comparison of the figures indicated that a greater average difference in course grades of the control and trained groups was accompanied by a greater number of common factors in all of the classifications, some of the averages indicating amount of course similarity for those course combinations having positive practice effect being almost double those for the combinations with negative practice effects.

A further indication of this positive relationship was secured by a comparison of five averages rather than two. The seven positive practice effects were divided into two divisions, and the twelve negative ones into three divisions according to the magnitude of the practice effects. The practice-test course combinations in these divisions were:

Division I	M P, S W, W P (low negative practice effect)
Division II	E P, M W, S P
Division III	W S, P S, M S
Division IV	S M, W M, E S
Division V	M E, S E, E M, W E (high positive practice effect)

TABLE 7
SHOWING THE GRADE DIFFERENCES BETWEEN THE CONTROL AND TRAINED GROUPS AND THE NUMBER OF COMMON FACTORS FOR THE PRACTICE-TEST COURSE COMBINATIONS ARRANGED ACCORDING TO THE AMOUNT OF THE GRADE DIFFERENCES

<i>Practice-Test Course Combination</i>	<i>Diff. in Grades</i>	<i>Tools</i>	<i>Materials</i>	<i>Total</i>	<i>Operations</i>	<i>Total¹</i>	<i>Product²</i>	<i>Rank Order Total</i>
ME	+.45	2	0	2	1	3	1	5
SE	+.40	7	1	8	3	11	1	9
EM	+.28	2	0	2	1	3	1	5
WE	+.21	6	3	9	3	12	1	10
SM	+.17	5	1	6	3	9	0	7
WM	+.15	3	0	3	2	5	0	3
ES	+.03	7	1	8	3	11	1	9
Average		4.57	0.86	5.29	2.29	7.71	0.71	6.88
WS	-.03	7	0	7	4	11	1	8
PE	-.11*	0	0	0	1	1	1	6
PS	-.12	0	1	1	0	1	0	2
MS	-.13	5	1	6	3	9	0	7
EP	-.22	0	0	0	1	1	1	6
MW	-.29	3	0	3	2	5	0	3
SP	-.30	0	1	1	0	1	0	2
MP	-.30	0	1	1	0	1	1	4
SW	-.30	7	0	7	4	11	1	8
PW	-.36*	0	0	0	0	0	0	1
EW	-.55*	6	3	9	3	12	1	10
WP	-1.09	0	0	0	0	0	0	1
Average		2.33	0.50	2.84	1.50	4.08	0.50	4.83

NOTE: *This difference may be explained by a difference in the same direction in the mechanical ability test battery scores.
¹Tools and Materials.
²Tools, Materials and Operations.

Table 8 contains the averages of the number of common factors possessed by the combinations for each division. The trends of these averages indicated definitely a correspondence between the amount of practice effect and the number of common factors, and substantiated the conclusions made above when only two divisions were used. The correspondence of the figures with the Rank Order Total was almost perfect. Operations showed better correspondence than tools or materials. While the results are conclusive as far as tendencies were concerned, it could not be said that in every case the more tools, materials, operations, etc., common to two courses, the greater the practice effect.

Results. Method III

In Table 9 are shown the percentages of the number of times each practice-test course combination showed a higher practice effect standing than another combination and also

TABLE 8
AVERAGE NUMBER OF COMMON FACTORS IN FIVE PRACTICE-TEST COURSE COMBINATION DIVISIONS

<i>Common Factor Classifications</i>	<i>Divisions of Practice Effect</i>				
	I (low)	II	III	IV	V (high)
Similar Operations	1.75	1.00	2.33	2.66	2.00
Similar Tools	2.33	1.00	4.00	5.00	4.25
Similar Materials	.33	.00	.66	.66	1.00
Similar Product	.33	.33	.33	.33	1.00
Tools and Materials	2.33	1.00	4.67	5.67	5.25
Tools, Materials and Operations	4.00	2.33	5.66	8.33	7.25
Rank Order Total	4.33	3.66	5.66	6.33	7.25

TABLE 9
THE PERCENTAGE OF TIMES THE PRACTICE-TEST COURSE COMBINATION WHICH HAD MORE PRACTICE EFFECT THAN ANOTHER HAD MORE, THE SAME, AND FEWER NUMBER OF COMMON FACTORS

	<i>More Factors in Common</i>	<i>Fewer Factors in Common</i>	<i>Same No. Factors in Common</i>
Similar Operations	48.54%	30.41%	21.04%
Similar Product	34.52	16.95	48.53
Similar Tools	52.63	29.29	18.13
Similar Materials	31.58	25.25	43.27
Tools and Materials	56.73	33.33	19.94
Tools, Materials and Operations	56.14	30.99	12.28
Rank Order Total	61.18	33.33	5.29

had more, the same, and less common factors for each classification. These percentages indicated that the more factors common to two courses, the greater the practice effect. As in the previous cases, the results were not conclusive enough to make it possible to state definitely a step by step parallelism. There was indicated, however, a fair degree of correspondence between the amount of practice effect and the number of common characteristics.

Results. Method IV

In the first method the data were arranged with reference to the same practice course and different test courses. In this method the procedure was reversed. The average amount of difference between the control and the trained groups in course grades was secured for each of the five test courses, in four practice-test course combinations.

<i>Test Course</i>	<i>Ave. Amount of Benefit Received from Practice Courses</i>
Electricity	+.24
Mechanical Drawing	+.20
Sheet Metal	— .06
Woodwork	— .38
Printing	— .48

If there were a perfect relationship between common factors and practice effects, then, the courses which had the most factors in common with the other four courses should obtain more benefits from practice than the courses with less factors in common. Table 10 contains the results. The courses are arranged in order from left to right according to the magnitude of the average practice benefit.

The results were not so conclusive under this arrangement. The position of the Printing course with the least benefit from practice in the other courses had the same relative standing in most of the common factor classifications. Woodwork, next to Printing in the amount of benefit, should have had a much higher position according to the number of factors common with other courses. The averages under Rank Order Total corresponded quite closely (one exception) with the ranking of the courses according to the amount of practice effect. The results showed that courses with more common factors received greater benefit from practice in each other than did those with less factors in common.

It was possible to secure the percentages of times that a

course received a greater practice benefit, also had more, the same and less common factors than another. These percentages, given in Table 11, substantiated the conclusion made previously, that the more common factors, the greater the practice effect.

These four methods of testing the data showed that there was a direct and positive relationship between the number of common factors in two courses and the amount of practice effect. A total (Rank Order Total), secured by adding ranks in operations, tools, materials, and products gave, for all five common factor classifications, the highest relationships.

TABLE 10
AVERAGE NUMBER OF COMMON FACTORS EACH COURSE HAD WITH THE OTHER COURSES. (ARRANGED ACCORDING TO THE AMOUNT OF PRACTICE BENEFIT—PRINTING LEAST)

	<i>Printing</i>	<i>Woodwork</i>	<i>Sheet Metal</i>	<i>Mechanical Drawing</i>	<i>Electricity</i>
Similar Operations	.25	2.25	2.59	2.00	2.00
Similar Tools	.00	4.00	4.75	3.33	3.75
Similar Materials	.25	.75	.75	.33	1.00
Similar Products	.50	.50	.50	.25	1.00
Tools and Materials	.25	4.75	5.50	3.66	4.50
Tools, Materials and Operations	.75	7.25	8.00	5.66	6.75
Rank Order Total	3.25	5.50	6.50	5.00	7.50

TABLE 11
CONTAINS THE PERCENTAGES OF TIMES A COURSE WHICH OBTAINED A HIGHER AVERAGE PRACTICE BENEFIT THAN ANOTHER COURSE HAD MORE, LESS, AND THE SAME NUMBER OF COMMON FACTORS WITH ITS PRACTICE COURSE.

	<i>More Factors in Common</i>	<i>Less Factors in Common</i>	<i>Same No. Factors in Common</i>
Similar Operations	70	20	10
Similar Tools	60	40	00
Similar Materials	70	20	10
Similar Products	40	30	30
Tools and Materials	60	40	00
Tools, Materials, Operations	60	40	00
Rank Order Total	80	20	00

Results. Method V

There was one other possible line of study which involved the testing of the following hypothesis: If two courses are similar to one another, each of them should have approximately the same effect on a third course. Since the best cor-

respondence between the amount of practice effect and the number of common factors was shown to be that with Rank Order Total, a ranking of the ten pairs of courses, according to it, was used in this connection. The ranking was as follows:

<i>Courses</i>	<i>Rank</i>	
W E	10	(most common factors)
S E	9	
S M	8	
E M	7	
W S	6	
W M	5	
E P	4	
M P	3	
S P	2	
W P	1	(least common factors)

The amount of practice effect which W, for example, exerted on a third course, M, was subtracted from the amount of practice effect exerted by E (the second named course) on the third course M. This figure gave an indication of the similarity in the amount of practice effect. An average of these differences for those pairs of courses having ranks 6 to 10 (more common factors), .082, was compared to the average for those pairs having ranks 1 to 5, .1175. The difference indicated that two courses which have more factors in common tended to have more similar effects on a third course than that of two courses with less factors in common.

Conclusions

1. There was definite indication of a positive correspondence between the amount of practice effect and the number of points of similarity in the practice and its test course.
2. Any measure of similarity (operations, tools, materials, or products) showed a positive relationship with the amount of practice effect.
3. A measure of similarity, composed of tools, materials, products and operations in equal proportion, gave the highest relationship to the amount of practice effect.
4. Of the single classifications, operations gave the best relationship.
5. Two courses which were more alike in terms of tools, materials, products, and operations tended to affect a third course more nearly the same than did two courses with less characteristics in common.

SECTION V

THE EFFECT OF LONGER PERIODS OF SHOP TRAINING

It was impossible in this investigation to measure the cumulative effects of practice on the sub-groups because the method of rotation through the courses was by individual rather than by group. It was possible, however, to do this for the entire group of subjects, irrespective of the particular sequences in which the students took the courses. That is, an index of the average quality of work of all students was obtained by averaging the grades for each, without regard to the courses taken. The method of grading the course products made this feasible, as all scores were placed in terms of deviations from the standard and converted into sigma deviations from the average of the entire group. For example, a grade of 5.00 sigma in Sheet Metal was equivalent to a grade of 5.00 sigma in Woodwork.

Table 12 contains the number of students in each ten weeks' training period, the average course grades and the mechanical ability test average scores. The figures indicated that there was no regular progression of improvement as the training period lengthened.

TABLE 12
AVERAGE GRADES AND THE MECHANICAL ABILITY TEST BATTERY AVERAGE
SCORES FOR THE STUDENTS AFTER 10, 20, 30, 40, AND 50 WEEKS OF
TRAINING

<i>Training Period</i>	<i>No. of Cases</i>	<i>Ave. Grades*</i>	<i>Mechanical Ability Ave. Scores</i>
10 weeks	146	5.01	2125
20 weeks	149	4.93	2125
30 weeks	142	5.23	1986
40 weeks	118	4.98	2044
50 weeks	13	5.60	2076

NOTE: * Low score indicates high accomplishment.

Table 13 shows the difference, the significance of the difference in the course grades, the difference in the mechanical ability test scores between the training periods of 10 and 20 weeks, 10 and 30 weeks, 10 and 40 weeks, 20 and 30 weeks, etc. The differences were not statistically significant (none over 2.00 sigma). A consideration of the direction of the differ-

TABLE 13

CONTAINS THE DIFFERENCE AND THE SIGNIFICANCE OF THE DIFFERENCE IN THE COURSE GRADES AND THE DIFFERENCE IN MECHANICAL ABILITY SCORES BETWEEN THE TRAINING PERIODS

<i>Training Periods Compared</i>	<i>Course Diff.</i>	<i>Grades Sig. of Diff.</i>	<i>Mechanical Diff.</i>	<i>Ability Sig. of Diff.</i>
10 and 20 weeks	+.08	.35	0	0
10 and 30 weeks	— .22	.80	— 139	.15
10 and 40 weeks	+.03	.09	— 81	.09
10 and 50 weeks	— .59	1.56	— 49	.05
20 and 30 weeks	— .30	1.22	— 139	.15
20 and 40 weeks	— .05	.16	— 81	.09
20 and 50 weeks	— .67	1.88	— 49	.05
30 and 40 weeks	+.25	.71	+ .58	.06
30 and 50 weeks	— .37	.95	+ .90	.10
40 and 50 weeks	— .62	1.43	+ .36	.04

ences indicated that there was evidence of negative practice effect. The number of minus signs, indicating negative influence or a decrease in the average grade the longer the training period, was greater than the number of plus signs, indicating a higher average grade after longer training. In fact seven of the ten indices were in the negative direction, four of which had an index of significance of over 1.00. None of the three plus differences attained this figure. The data indicated that higher quality of work did not result from longer training.

For several reasons, the size of the group decreased as the training period lengthened. In order to determine whether this change was the cause of the fluctuations in average grade, the average score in the mechanical ability test battery for each period was secured, the data for which are given in the second vertical division of Table 13. The differences showed that the group was very similar in this regard throughout the course of the experiment, since in no case was the difference in mechanical ability larger than one-fifth the sigma of the distribution. Taking the three plus tendencies in practice effect, there was no change in mechanical ability for one comparison, a slight increase for another, and a slight decrease for the third. In the seven negative practice tendencies, five were accompanied by a small decrease in mechanical ability and two were accompanied by an increase.

The explanation of the occurrence of negative transfer in this investigation may be found in one or more of the follow-

ing causes. (1) Habits of work established in one kind of training may be of such a nature as to block the forming of correct habits in another course. For example, the method of laying out dimensions in Woodwork in terms of linear measurements is very different from the method of laying out dimensions according to angles and circumferences in Sheet Metal. (2) It may be necessary to use the same tool differently in two shops. The methods of using a hammer in Woodwork are different from the methods of using the same tool in Sheet Metal. (3) Tools used for the same purpose in two courses may differ in certain characteristics. The rulers used in Sheet Metal are scaled differently from those used in Woodwork, Mechanical Drawing, and Electricity. (4) The amount of interest the students take in the different shops may not be constant. Constant interest was in evidence throughout the course of the investigation in all the courses except Printing. This may account for the negative practice effects of other courses on Printing.

Conclusions

1. There was no regular progression of improvement in shop work as the period of practice lengthened.
2. There was some indication that the effects of practice were negative in character; that is, as the period increased in length, the work of the students became inferior to that produced earlier.
3. These results were not caused by a change in the mechanical ability of the group as the period lengthened.

SECTION VI

SUMMARY

1. *Procedure*

The effects of practice were studied in five school shop courses. Comparisons of average grades of groups with certain kinds of training and groups without such training were made. Course grades were based on objective measurements of performance. The equality of the compared groups was determined by average scores on a battery of three mechanical ability tests. Correlations of shop proficiency with standing in chronological age and scores on an intelligence test indicated that these last two factors had little bearing on shop work. The differences in mechanical ability, as measured by the test scores, were taken into consideration in determining whether an indicated practice effect was caused by group dissimilarity.

Measures of course similarity were secured on several characteristics, such as, operations, tools, materials, products, and certain combinations of these. In all cases these measures were the result of observation and not of personal judgment. A comparison of the number of similar points between the courses in the various practice-test course combinations and the amount of practice effect was made.

Finally indices of the cumulative effect of practice in ten, twenty, thirty, and forty weeks' divisions were obtained. This procedure was feasible because of the manner in which the objective grades had been converted into comparable units.

2. *Results*

1. The reliability of the indices of practice effect was low. In these courses and under the conditions of the experiment very little transfer, either of a positive or negative nature, occurred.

2. The indicated practice effects were not always favorable. In fact, twelve of the nineteen situations studied showed that practice had an unfavorable effect.

3. The differences indicating practice effect were due to variations in practice and not to group dissimilarity, except in three cases.

4. Previous practice of the kind studied here was unfavorable to work in Printing, Woodwork, and Sheet Metal.

5. Previous practice was beneficial for work in Mechanical Drawing and Electricity.

6. A tentative shop curriculum based on the indicated practice effects should have this arrangement:

Course I	Printing
Course II	Woodwork
Course III	Sheet Metal
Course IV	Electricity
Course V	Mechanical Drawing

7. A definite and positive relationship existed between the amount of practice effect and the number of points of similarity contained in the practice and test courses.

8. A measure of similarity, composed of tools, materials, products, and operations in equal proportion, showed the best relationship with the amount of practice effect.

9. Of the individual classifications, similar operations gave the best measure of relationship with amount of practice effect.

10. A pronounced tendency for two courses to have the same general effect on each other was noted.

11. Two courses which were more alike in terms of tools, materials, products, and operations tended to have the same kind of effect on a third course.

12. Continued practice of a varied nature in shop work was unfavorable to the quality of work performed.

13. In regard to the problem of transfer of training, the results indicated:

- A. That the size of a practice effect is proportional to the number of similar factors in two situations.
- B. That in these courses and under similar school conditions, transfer of training is more apt to be negative in character than positive. A conclusion that negative transfer occurs more often than positive transfer is, however, a result of the small number of courses studied. It may mean that the five courses were not homogeneous and should not be considered as such because of their inclusion in the same shop curriculum. A conclusion concerning negative transfer within a small number of courses cannot be *generalized*.

- C. That the correspondence of the amount of positive practice effect and the number of common factors appeared in any of the possible ways of measuring similarity—tools, operations, materials, or products.

VITA

L. Dewey Anderson.

Born, July 8, 1898, Laramie, Wyoming.

A.B., University of Wyoming, 1920.

M.A., Carnegie Institute of Technology, 1921.

Fellow, Bureau of Personnel Research, Carnegie Institute of Technology, 1921-22.

Chief Investigator, Mechanical Ability Project, Committee of Human Migrations, National Research Council, 1924-27.

Research Assistant Professor, Department of Psychology, University of Minnesota, 1924-27.

Psychologist, Bureau of Educational Experiments, 1927-29.